Politecnico di Torino
1859

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (35thcycle)

Multi-Modal Learning for Cross-Domain Analysis of Egocentric Action and Object Recognition

By

Mirco Planamente

******

Supervisor(s):
Prof. Barbara Caputo, Supervisor
Prof. Andrea Bottino, Co-Supervisor

Doctoral Examination Committee:
Prof. Ronald Poppe, Referee, University of Utrecht
Prof. Oswald Lanz, Referee, University of Bolzano
Prof. Paolo Garza, Polytechnic of Turin
Dr. Pietro Morerio, Italian Institute of Technology
Prof. Marco Torchiano, Polytechnic of Turin

Politecnico di Torino
2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Mirco Planamente
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

I would like to dedicate this thesis to my cherished family, supportive friends, and wonderful girlfriend for their unwavering encouragement and inspiration throughout my academic journey.

Abstract

The field of artificial intelligence has witnessed remarkable progress in the last few years, but there still exists a disparity between the capabilities of intelligent systems and those of humans in terms of accuracy and efficiency. Humans possess the ability to perceive and interact with the world using multiple senses, enabling a deeper understanding beyond what can be achieved with a single 2D representation. Motivated by this, the focus of this thesis is on multi-modal learning to enhance model performance in terms of accuracy, robustness, and adaptability.

The thesis explores two interrelated research domains: Object Recognition (OR) and Egocentric Action Recognition (EAR). These domains present specific challenges that conventional uni-modal approaches, usually image-based models, struggle to address. The thesis explores two interconnected research domains: Object Recognition (OR) and Egocentric Action Recognition (EAR). These domains pose specific challenges that conventional uni-modal approaches, primarily image-based models, struggle to address. Image-based models tend to exhibit a bias toward texture and color information while encoding a limited amount of geometric and motion cues. Additionally, they are sensitive to variations in lighting and other environmental conditions, which compromises their generalization capability. Consequently, their performance may suffer when applied to diverse real-world scenarios. This research explores the integration of both visual and non-visual modalities, with the aim of leveraging their complementary characteristics and capturing the intricate nature of real-world information. The thesis proposes techniques to enhance model capabilities, enabling robustness and adaptability to different environments and tasks. Our research in multi-modal learning also includes the exploration of alternative modalities derived from innovative devices that go beyond traditional sensors, like event-based cameras.

Overall, this thesis contributes to the advancement of multi-modal learning for cross-domain analysis of EAR and OR, aiming to enhance the capabilities of AI systems and reduce the gap between human and machine perception.

Keywords: deep learning, multi-modal learning, domain adaptation, domain generalization, egocentric vision, event-based vision

Contents

List of Figures **x**

List of Tables **xix**

1 Introduction **1**

1.1 Recognize Objects using Multi-Modal data 4

1.2 Recognize Actions in First Person videos 6

1.3 Contributions 8

1.4 Outline 9

1.5 Publications 10

1.5.1 Technical Report 13

2 Background and Related Work **14**

2.1 Learning to recognize: From Object to Action 15

2.1.1 Robotics Vision: Object Recognition from RGB-D data . . 15

2.1.2 Action Recognition 17

2.1.3 First Person Action Recognition 19

2.2 Learning to see across Domains and Modalities 20

2.2.1 Problem Formulation 21

2.2.2 Unsupervised Domain Adaptation 22

2.2.3 Domain Generalization 25

2.2.4 Multi-Modal Adaptation 26

2.3 Learning to see from Event data 27

2.3.1 Grid-Like Event Representations 31

2.3.2 Data Scarcity Challenge in Event-Based Vision 33

2.4 Dataset 34

2.4.1 Robotics Vision Datasets 35

2.4.2 Action Recognition Datasets 40

3 Multi-Modal Learning for Robotics Vision: Object Recognition 45

3.1 Multi-Modal Fusion – RCFusion 46

3.1.1 Recurrent Convolutional Fusion 49

3.1.2 Experiments 53

3.1.3 Discussion and Conclusion 59

3.2 Multi-Modal Alignment – RR 61

3.2.1 Dataset 64

3.2.2 Method 66

3.2.3 Experiments 71

3.2.4 Conclusion 76

3.3 From Pixels to Events – DA4E 78

3.3.1 DA4Event 80

3.3.2 MV-DA4Event: a Multi-View Approach 82

3.3.3 Experiments 84

3.3.4 Conclusions 92

4 Multi-Modal Learning for Egocentric Vision: First Person Action Recognition 93

4.1 Single-stream Multi-Modal Fusion – SparNet 94

4.1.1 Architecture Overview and Details 96

4.1.2 Experiments 101

4.1.3 Conclusions 107

4.2 Multi-Modal Alignment – RNA 109

4.2.1 Intuition and Motivation 112

4.2.2 Relative Norm Alignment Loss 113

4.2.3 Learning objective in UDA 117

4.2.4 Experiments 118

4.2.5 Experiments on EK-100 119

4.2.6 Experiments on EK-55 130

4.2.7 Experiments on UCF-HMDB 133

4.2.8 Experiments on ROD 135

4.2.9 Experiments on CogBeacon 136

4.2.10 Limitations 138

4.2.11 Conclusion 138

4.3 From Pixels to Events – E^2(GO)MOTION 140

4.3.1 N-EPIC-Kitchens 142

4.3.2 Event-Based Vision Data 143

4.3.3 Learning from Motion 146

4.3.4 Experiments 149

4.3.5 Conclusion 156

5 Conclusion 159

5.1 Summary 160

5.2 Open Issues 161

References 167

Appendix A Code Repositories 207

Appendix B Project Contributions & Computational Resources Support 209

List of Figures

1.1 These samples depict object and action frames extracted from the RGB-D Object dataset (ROD) [1] and EPIC-Kitchens-55 (EK55) dataset [2], showcasing their complementarity in terms of modalities. The ROD dataset includes depth images that provide additional geometric information, while the EK55 dataset incorporates optical flow information that captures motion details. These samples showcase the diverse types of information encoded by these modalities, highlighting their potential to complement standard RGB images and provide a deeper understanding of the objects or actions. 2

1.2 These samples showcase object and action frames obtained from the RGB-D Object dataset (ROD) [1] and its synthetic counterpart synROD [3], and two different kitchens of the EPIC-Kitchens-55 (EK55) dataset [2]. On the left side, the samples highlight the synthetic-to-real shift problem, illustrating the disparity in data distribution between the synthetic data generated using Blender [4] and the real data captured with a camera. On the right side, the samples demonstrate the environment bias, depicting the significant variation in appearance between data recorded in different kitchen environments. 3

2.1 Samples from the RGB-D Object dataset [1], showcasing two modalities: RGB (top) and depth images (bottom). The data has been processed according to the procedure outlined in [5]. The depth information is represented using surface normals, employing the best non-learned colorization method currently available. 35

2.2 Samples from the RGB-D image dataset [6] used in this study. The dataset comprises images of workshop tools, including clamps and screwdrivers. 36

2.3 Samples from the Object Clutter Indoor (OCID) dataset, showcasing two modalities: RGB (top) and depth images (bottom). The data has been processed according to the procedure outlined in [5]. The depth information is represented using surface normals, employing the best non-learned colorization method currently available. 36

2.4 Samples from the synthetic dataset synROD [3] and its real counterpart RGB-D Object dataset (ROD) [1]. The top two lines depict RGB and depth images from synROD, while the bottom two lines showcase the corresponding RGB and depth images from ROD, emphasizing the domain shift from synthetic to real data. The data has been processed according to the procedure outlined in [5]. The depth information is represented using surface normals, employing the best non-learned colorization method currently available. 37

2.5 Samples from the Homebrewed dataset [7], synthetic (synHB) and real counterpart (realHB). The top two lines display RGB and depth images from synHB, while the bottom two lines show the corresponding RGB and depth images from realHB. The data has been processed following the procedure described in [5]. Depth information is encoded using surface normals, utilizing the state-of-the-art non-learned colorization method for enhanced visualization. 38

2.6 Real and simulated events (voxel grid [8]) on a Caltech101 sample. . 39

2.7 Samples from the ROD [1] dataset (a)-(b), and from the synthetic version synROD [3] (c)-(d). Event sequences are displayed using a voxel-grid [8] representation. 39

2.8 Samples from the Georgia Tech Egocentric Activities (GTEA) dataset [9]. The videos are captured using a GoPro camera mounted on a baseball cap, which is worn by the subject. The camera is positioned to record the visual field in front of the subject's eyes, providing a first-person perspective. 40

2.9 Samples from the extended GTEA Gaze+ dataset [10]. The dataset comprises 28 hours of cooking activities including 106 fine-grained action classes with an average duration of 3.2 seconds. 41

2.10 Samples from the First-Person Hand Action (FPHA) benchmark dataset [11]. Frames capturing hand actions recorded in three distinct scenarios are shown. 41

2.11 Samples from the EPIC-Kitchens-55 (EK55) dataset [2]. This dataset is a comprehensive and diverse collection of egocentric videos of fine-grained actions captured in the kitchens of 32 participants from 10 different countries. The videos are recorded using head-mounted cameras and are of Full HD quality. 42

2.12 The first row showcases samples from the EPIC-Kitchen-55 (EK55) dataset [12], while the second row features samples from the extended version, EPIC-Kitchen-100 (EK100) dataset [13]. EK100 expands upon EK55 with 100 hours of video, 20 million frames, and 90,000 actions across 700 videos. In our study, we focus on the challenging task of unsupervised domain adaptation, which involves adapting from a labeled source domain (videos recorded in 2018) to an unlabeled target domain (newly collected videos). The domain shift arises from changes in location, hardware, and temporal offsets. 43

2.13 Samples from the HMDB51 and UCF101 datasets [14, 15] . The first row showcases examples of different actions from the HMDB51 dataset, while the second row displays the corresponding action samples from the UCF101 dataset. 44

3.1 High-level scheme of RCFusion. The blue boxes are instantiated with convolutional neural networks and the thick arrows represent multiple feature vectors extracted from different layers of a CNN. . 47

3.2 The architecture of Recurrent Convolutional Fusion (RCFusion). It consists of two parallel streams of Convolutional Neural Networks (CNNs) that process RGB and depth images respectively. The features extracted from the corresponding hidden layers of the two CNNs are then projected into a shared representation space, concatenated, and sequentially fed into a Recurrent Neural Network (RNN) for synthesis. The final multi-modal features produced by the RNN are then utilized by a classifier to predict the label of the input data. 49

3.3 Projection block. It converts the input feature f_i^* into a projected feature p_i^*. The block consists of a convolutional layer $conv(k \times k) \times D$ with D filters of size $(k \times k)$, followed by a batch normalization layer (BN) and an activation layer with ReLU non-linearity ($ReLU$). 51

3.4 Per class accuracy (%) of RCFusion on RGB-D Object Dataset [16]. 54

3.5 The figure presents a t-SNE visualization of features extracted from three modalities (RGB, Depth, and RGB-D) in object recognition. . 57

3.6 Object crops with their respective instance labels obtained from the Object Cluttered Indoor Dataset [17]. The figure shows a selection of examples from the dataset, highlighting the diversity of objects and scenes included in the dataset. 58

3.7 Q: “By how much should the RGB image (top) be rotated to align with the depth image (bottom)?" A: “90°". This question describes the self-supervised task of predicting the relative rotation between the RGB and depth image (shown with surface normal colorization [18]) of a sample after they have been independently rotated. 62

3.8 Overview of our method for RGB-D domain adaptation, which uses a convolutional neural network (blue squares) consisting of a two-stream feature extractor E and two network heads. The main head M is trained for object recognition using the labeled source data (red arrow), while the pretext head P is trained using both source and target data (orange+red arrow), with independent rotation of RGB and depth images 63

3.9 Examples of rendered scenes from synROD dataset. For each, we showcase the RGB, raw depth and segmentation mask image, with increasing level of clutter from top to bottom. 65

3.10 Examples images from PACS [19] (top row) and HomebrewedDB [7] (bottom row) that are rotated by 0°, 90°, 180°, and 270°. While it is possible to infer the rotation of the PACS samples based on the contextual background and prior knowledge of the subject matter, the same cannot be said for the HomebrewedDB samples. This observation highlights the inadequacy of predicting image rotation by analyzing each image in isolation, as proposed in [20]. Hence, this approach is deemed ill-posed. 69

3.11 All the possible combinations of RGB and depth rotation for a given relative rotation $\{0°, 90°, 180°, 270°\}$. 70

3.12 The figure presents a visualization of the significant pixels for predicting the relative rotation. The input of the network is an RGB-D image, labeled as “original." The saliency map of the input is generated using the last layer of the feature extractors E^c and E^d, denoted as “guided backprop" [21]. The “binary backprop" is a binarized version of the “guided backprop" map that highlights peak values in white to aid in visualization. The depth image is utilized with surface normal colorization, following the technique proposed by [18]. 74

3.13 t-SNE [22] visualization of the HomebrewedDB [7]. Features extracted from the last hidden layer of the main head M are used for the plot. Red dots: source samples; blue dots: target samples. When adapting the two domains with our method (right), the two distributions align much better compared to the non-adapted case (left). 75

3.14 *How can we bridge the Sim-to-Real gap in event-based cameras?* DA4Events leverages unsupervised domain adaptation techniques at the feature level to mitigate this issue. *How else simulated events can be used?* Our proposed approach involves utilizing event data in real-world scenarios, thereby enhancing network robustness through the complementary integration of RGB data. 79

3.15 Real and simulated events (voxel grid [8]) on a Caltech101 sample. . 81

3.16 Overview of the DA4Event pipeline. **Top** shows the process of extracting an event representation, using voxel grids [8] and three views as an example.**Bottom** details the proposed multi-view architecture (MV-DA4E). During training, two unpaired random batches from source and target domains are sampled and processed separately. When the multi-view approach is not used (DA4E), event representations are fed as a single multi-channel tensor to the feature extractor F, and multi-view pooling is removed. Notice that only source (labeled) data are fed to the classifier G, while both target and source data are fed to the DABlock. 82

3.17 Difference in terms of performance based on percentage (%) of target data used during training, obtained with constant threshold $C = 0.06$. 88

3.18 t-SNE visualization of N-Caltech [23] features from the last hidden layer of the main classifier. Red dots: source samples; blue dots: target samples. When adapting the two domains with the proposed DA4E (b), the two distributions align much better compared to the non-adapted case (a). 88

3.19 Grad-CAM [24] visualizations on several real N-Caltech101 samples. In each triplet we show the input event representations (voxel grid [8]), the activation maps when the network is trained on simulated data only, and those obtained by training with MV-DA4E. 89

4.1 By leveraging a self-supervised motion prediction task during training, at test time, SparNet can jointly exploit motion and appearance information using a single RGB stream. The result is a leaner architecture that better focuses on the relevant elements for action recognition in egocentric vision, as can be seen from the comparison between the Class Activation Maps when the auxiliary task is used (lower path) or not (upper path) 95

4.2 SparNet architecture. The action recognition block computes image embeddings and solves the classification task. The motion prediction block injects into the backbone motion information that results in a richer embedding, jointly encoding motion and appearance information. We compensate for egomotion effects with the use of stabilized dense optical flow and an IDT module [25]. The OF quantizer block discretizes the input motion vectors into a finite number of classes according to the values of the n_d and n_m parameters (respectively, 8 and 3 in this example, for a total of 17 classes). 96

4.3 Top-3 results of the 2019 [26] and 2020 [27] EK challenges, when testing on “Seen" and “Unseen" kitchens. 110

4.4 Overview of Relative Norm Alignment (RNA) loss for RGB and audio modalities. Given visual and audio input from both source and target domains, we perform an alignment at feature level by rebalancing (i) the mean feature norms of visual and audio modalities (*cross-modal alignment*, $\mathcal{L}^{g}_{RNA}$), (ii) per-class mean feature norms of visual and audio modalities (*per-class alignment*, $\mathcal{L}^{c}_{RNA}$) and (iii) mean feature norms of source and target features independently for each modality (*cross-domain alignment*, $\mathcal{L}^{mod}_{RNA}$). 111

4.5 Labeled source and unlabeled target samples from the modalities u (e.g., visual) and v (e.g., audio) are fed to the respective feature extractors. $\mathcal{L}_{RNA}$ aims to balance the relative feature norms of the two modalities, through a combination of the (domain-specific) cross-modal components ($\mathcal{L}^{g}_{RNA}$ and $\mathcal{L}^{c}_{RNA}$) and the cross-domain ones ($\mathcal{L}^{mod}_{RNA}$) in each u and v modality. In DG, only the components computed on the source are used. 114

4.6 *Individual effects* of the different components of $\mathcal{L}_{RNA}$ on the feature norms. Each diagram shows the norm per class of a single modality and a single domain (the color coding is the same as in Fig. 4.5). **First row:** $\mathcal{L}^{g}_{RNA}$ minimizes the overall average of the feature norms (larger bar on the right) of the different modalities (u and v for either source or target). **Second row:** $\mathcal{L}^{c}_{RNA}$ achieves the same goal at the individual class level (left: unbalanced norms, right: balanced norms thereafter). **Third row:** $\mathcal{L}^{mod}_{RNA}$ minimizes the difference between global and per class feature norms of the same modality across different domains. Here, diagrams represent the class and average norms for the same modalities (either u or v) before (left, unbalanced between domains) and after (right, balanced) the application of $\mathcal{L}^{mod}_{RNA}$. 116

4.7 Verb feature norms across different modalities and settings (DG and UDA). Light (🟦 🟩 🟧) and dark colors (🟦 🟩 🟧) denote source and target validation domains, respectively. **(a)** In the *"Source Only"* setting, different modalities and domains result in unbalanced feature norms. **(b)** $\mathcal{L}_{RNA}$ in DG improves the alignment between different modalities, but leaves a gap between the source and target domains. **(c)** Finally, the contribution of $\mathcal{L}^{mod}$ in $\mathcal{L}_{RNA}$ reduces this gap in UDA, resulting in more consistent feature norms across different modalities and domains. 120

4.8 Feature norms of the top 10 most and least common classes from the target validation split of EPIC-Kitchen-100[13]. While $\mathcal{L}^{g}_{RNA}$ improves the alignment of different modalities, there is still an imbalance between classes. The addition of the per-class variant of RNA greatly improves this alignment, resulting in more uniform feature norms across different classes. 121

4.9 Comparison of the feature norms before (top) and after (bottom) application of $\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$. The dots represent the samples in the validation dataset. The color bar on the right represents increasing density values. The original features, i.e. *"Source Only"*, show a wide range of values and an irregular shape, reflecting the misalignment between the features norms of the two modalities. The RNA loss re-balances the two, as evidenced by the more globular distribution while also shifting the average norms towards higher values. 124

4.10 **N-EPIC-Kitchens**: the first event-based dataset for egocentric action recognition. From RGB images, we generate a stream of events (bottom). Positive polarity is represented by red events, whereas blue events represent negative polarity. Events focus on motion, similarly to optical flow (top). With their low latency, high temporal resolution, and low-power consumption, event data are a perfect fit for egocentric action recognition. 141

4.11 N-EPIC-Kitchens *vs* existing event-based action classification datasets in the literature [28–32]. 143

4.12 RGB (top), optical flow (middle) and Voxel Grid representation (bottom) from the same action ("cut") on the three different kitchens (D1, D2, D3). 144

4.13 Illustration of the proposed E^2(GO)MO. The input $\mathbf{x}^E$ and $\mathbf{x}^F$ from the event and flow modality are passed to the feature extractors F^E and F^F respectively. Information from the pre-trained teacher stream (frozen) F^F is distilled to the student stream F^E. The latter is trained with standard cross-entropy loss. 147

4.14 Accuracy *vs* time of RGB modality, E^2(GO)MO, estimated PWCNet optical flow and TV-L1 optical Flow on seen and unseen scenarios for one clip evalutation. 154

List of Tables

3.1 Top-1 accuracy (%) achieved by various object recognition methods on RGB-D Object Dataset [16]. The highest result is highlighted in **bold**, while other notable results are underlined. 55

3.2 Top-1 accuracy (%) achieved by various object recognition methods on the JHUIT-50 dataset [6]. The highest result is highlighted in **bold**, while other notable results are underlined. 55

3.3 Analysis of the most frequently misclassified classes in RGB, depth, and RGB-D modalities with respect to selected reference classes. . . 56

3.4 Top-1 accuracy (%) obtained by DECO [33] and variations of the RCFusion method on the Object Clutter Indoor Dataset [17]. The RCFusion method is modified in two ways, with “RCFusion - res5" referring to the variation that only uses features from the last residual layer (res5) for classification and “RCFusion - fc" referring to the variation that uses a fully connected layer instead of a recurrent neural network to combine RGB and depth features. The highest result is highlighted in **bold**, while other notable results are underlined. 59

3.5 Target Top-1 accuracy (%) of several methods for RGB-D domain adaptation on two datasets with synthetic-to-real shifts, synROD→ROD and synHB→realHB. The highest results are highlighted in **bold**. . . 73

3.6 Target Top-1 accuracy (%) of variations of our method for RGB-D domain adaptation on two synthetic-to-real shifts, synROD→ROD and synHB→realHB. The highest results are highlighted in **bold.** . . 74

3.7 Target Top-1 accuracy (%) of Unsupervised Domain Adaptation methods on N-Caltech101. The highest results are highlighted in **bold.** 87

3.8 Target Top-1 accuracy (%) of Unsupervised Domain Adaptation methods w.r.t. to methods that act on the contrast threshold C. The highest results are highlighted in **bold**. 90

3.9 Target Top-1 accuracy (%) of the event, RGB and depth modalities, both in single-modal and multi-modal (RGB+E). The highest result is highlighted in **bold**. 91

4.1 Top-1 accuracy (%) achieved by various state-of-the-art methods on GTEA-61, EGTEA+, and FPHA datasets. The highest results are highlighted in **bold**. 102

4.2 Top-1 accuracy (%) achieved by our method using a different number of representative frames and input features. The highest results are highlighted in **bold**. 102

4.3 Top-1 accuracy (%) achieved by variations of our method compared with others by various state-of-the-art methods on EGTEA+ (split 1, $N = 11$). The highest result is highlighted in **bold**. 104

4.4 Top-1 accuracy (%) achieve by applying the motion segmentation (MS) task on Ego-RNN [34] and LSTA [35]. 105

4.5 Top-1 accuracy (%) on the EPIC-Kitchens-55 validation set defined in [36]. The highest results are highlighted in **bold**. 107

4.6 Coefficient of variation for DG and UDA feature norms. CV$_S$, CV$_T$ and CV$_{S+T}$ are the CVs of the source, target and combined domain(s) respectively. For clarity, we also report the percentages of improvement with respect to the "Source Only" experiment. 122

4.7 Target Top-1 accuracy (%) achieved using various loss components. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics. The highest results are highlighted in **bold**, while other notable results are underlined. 125

4.8 Target Top-1 accuracy (%) on modality pairs on EPIC-Kitchens-100 [13]. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics. 126

4.9 Target Top-1 accuracy (%) obtained through ablation experiments by dropping a modality at test time. All configurations are trained on all input modalities. At inference time, we simulate the loss of a modality which results in a large performance drop. RNA helps the model avoid focusing too much on individual modalities and is able to mitigate the performance drop. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics. 127

4.10 Target Top-1 and Top-5 accuracy (%) on EPIC-Kitchens-100 [13]. Results are reported for the noun, verb and action metrics. Δ Acc. is the average Top1-accuracy improvement. †These experiments are trained using the cross entropy loss on both the fused logits as well as on the *per-modality* logits. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>. 129

4.11 Target Top-1 accuracy (%) on EPIC-Kitchen-55 [12], using the evaluation protocol from [37], divided by modalities. Results are grouped by the sampling strategy used for a fair comparison. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>. 132

4.12 Target Top-1 accuracy (%) on UCF-HMDB on RGB+Flow combination. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>. 134

4.13 Target Top-1 accuracy (%) on SynROD→ROD. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>. . 135

4.14 Target Top-1 accuracy (%) on CogBeacon. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>. . . . 137

4.15 Top-1 accuracy (%) achieved using I3D, TSN and TSM architectures depending on the number of channels for the event representation. The highest results are highlighted in **bold**. 150

4.16 Top-1 accuracy (%) of event w.r.t. RGB on both I3D and TSM. Results are shown on all shifts, i.e., $D_i \rightarrow D_j$ indicates we trained on D_i and tested on D_j, and D_i means we trained and test on the same. E^2(GO)-3D and E^2(GO)-2D improvements are shown w.r.t. to their respective baselines, where no architectural variations are performed. The highest results are highlighted in **bold** on both seen and unseen for each backbone. 151

4.17 Top-1 accuracy (%) of the event modality when used in combination to stardard RGB and optical flow. The highest results are highlighted in **bold**. 151

4.18 Top-1 accuracy (%) of E^2(GO)MO w.r.t. the baseline on events (TSM) and E^2(GO)-2D. We compare E^2(GO)MO with the same approach on RGB to validate the choice of combining event and flow. The highest results are highlighted in **bold**, while underlined the best multi-modal. 152

4.19 Top-1 accuracy (%) of RGB, Event and optical flow (TV-L1), along with their representation time, i.e., time to calculate the Voxel Grid for event, and extraction time for TV-L1 flow. The highest results are highlighted in **bold**. 155

Chapter 1

Introduction

Artificial intelligence (AI) is a rapidly evolving field of computer science that aims to develop algorithms and systems capable of emulating human intelligence. While significant progress has been made in AI, particularly in areas like image generation [38] and natural language processing [39], achieving true human-like understanding remains a complex challenge. Human intelligence encompasses various cognitive processes and is influenced by numerous factors, making it, even in the present day, a fascinating and intricate phenomenon and area to explore.

One remarkable aspect of human intelligence is our ability to perceive and interact with the world in a highly efficient and adaptive manner. A fundamental distinction between humans and machine lies in the process of learning itself. In fact, humans acquire their knowledge of the world through a multi-sensory perception, as emphasized in the research conducted by Bertelson et al. [40]. Starting from early childhood, our ability to recognize simple objects develops by integrating inputs from various sensory modalities. For example, vision provides an extensive range of information that goes beyond what a standard RGB camera can capture, while the sense of touch, that rapidly evolves into the capability of object manipulation, deepens our understanding of objects. This multi-sensory capacity extends to other modalities as well, enabling us to comprehend and interpret objects in a more comprehensive and contextual manner. In contrast, much of the progress in the field of AI has primarily focused on unimodal research. It began with the early successes in computer vision, where convolutional networks gained widespread adoption and large-scale datasets like ImageNet played a crucial role in advancing the field. The trend has continued with the emergence of transformer-based models and the use of web data for training, further pushing the boundaries of unimodal approaches [41]. This disparity between uni-modal and multi-modal learning approaches presents significant challenges in AI. This has motivated researchers to seek solutions to combine different modalities to enhance neural networks' ability to understand and represent the world in a more human-like manner.

The primary aim of this thesis is to explore multi-modal approaches for enhancing the learning capabilities of neural networks, introducing novel models, data fusion techniques, and domain adaptation strategies. In order to explore these concepts, the thesis will focus on two main tasks. The first task is Object Recognition (OR), which is a widely studied area in the field of robotics vision. OR serves as an interesting case to highlight the limitation of two-dimensional information captured by RGB cameras compared to the three-dimensional perception of humans. The second task is Egocentric Action Recognition (EAR) or First Person Action Recognition (FPAR), which is a more complex task. This involves recognizing actions based on egocentric data captured by a camera worn directly by the user. This unique setup offers an exceptional opportunity to gain deeper insights into the human learning process.

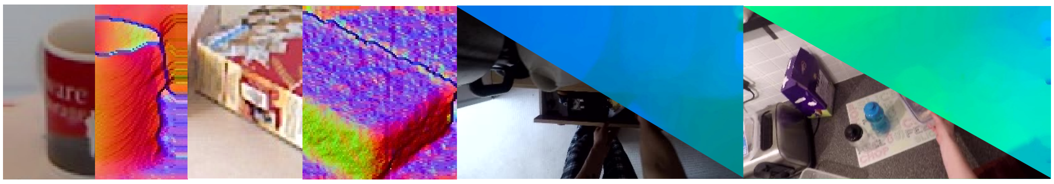


Fig. 1.1 These samples depict object and action frames extracted from the RGB-D Object dataset (ROD) [1] and EPIC-Kitchens-55 (EK55) dataset [2], showcasing their complementarity in terms of modalities. The ROD dataset includes depth images that provide additional geometric information, while the EK55 dataset incorporates optical flow information that captures motion details. These samples showcase the diverse types of information encoded by these modalities, highlighting their potential to complement standard RGB images and provide a deeper understanding of the objects or actions.

From the interconnected yet distinct contexts of OR and FPAR, three main research questions emerge. Firstly, we delve into the challenge of **effectively leveraging multiple modalities to overcome the limitation of uni-modal approaches**, which usually tend to overfit on appearance information and struggle to encode geometric or motion information (as shown in Figure 1.1) which are fundamental cues for different recognition tasks.

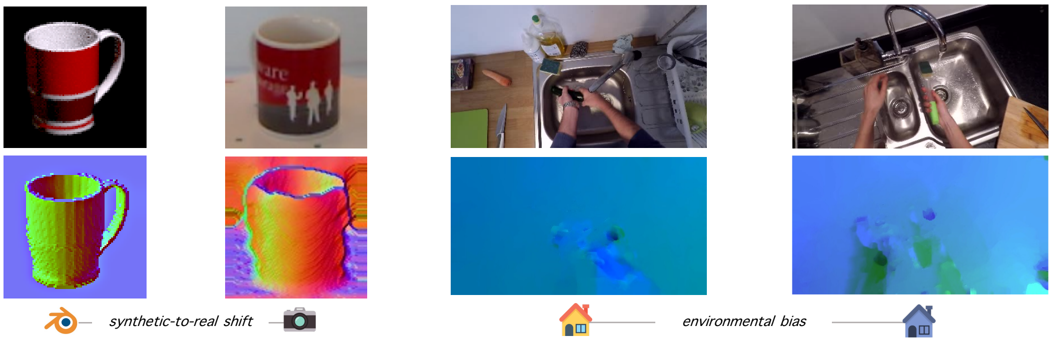


Fig. 1.2 These samples showcase object and action frames obtained from the RGB-D Object dataset (ROD) [1] and its synthetic counterpart synROD [3], and two different kitchens of the EPIC-Kitchens-55 (EK55) dataset [2]. On the left side, the samples highlight the synthetic-to-real shift problem, illustrating the disparity in data distribution between the synthetic data generated using Blender [4] and the real data captured with a camera. On the right side, the samples demonstrate the environment bias, depicting the significant variation in appearance between data recorded in different kitchen environments.

Secondly, despite the remarkable advancements in image-based recognition models, they still struggle to handle the inherent variations in data distribution, especially when the data originates from different domains. This phenomenon is commonly referred to as domain shift and it has a significant impact on the performance of recognition models. In Figure 1.2, two examples of domain shift are presented, specifically in the context of OR and FPAR. These challenges are referred to as the synthetic-to-real shift and environmental bias, respectively (more detailed explanations will be provided in Sections 1.1 and 1.2). To address this challenge, it is crucial to explore the research fields of Domain Generalization (DG) and Unsupervised Domain Adaptation (UDA), which are fundamental to effectively handling and minimizing the adverse effects of domain shift. In particular, in the context of multi-modal learning, different modalities are influenced by domain shifts in a unique manner, motivating our research **to investigate approaches that leverage the modalities to improve the generalization and adaptation abilities of existing models.**

Lastly, multi-modal learning extends beyond traditional modalities like RGB, audio, and depth. It encompasses **the exploration of alternative modalities derived from novel devices that go beyond conventional sensors**. By investigating and incorporating novel modalities into our research, we aim to broaden the scope of multi-modal learning and uncover its potential benefits in various domains. In the subsequent paragraphs, we will delve deeper into these research topics, providing the reader with a more detailed understanding of the origins of these research questions. We will also explore the distinct nature of these topics within the domains of object recognition (OR) and first-person activity recognition (FPAR).

1.1 Recognize Objects using Multi-Modal data

Human-built environments are, ultimately, collections of objects. Every daily activity requires understanding and operating a set of objects to accomplish a task. Robotic systems that aim to assist the user in his own environment need to possess the ability to recognize objects. In fact, OR is the foundation for higher-level tasks that rely on an accurate description of the visual scene. Despite the interesting results achieved in this context by operating on standard RGB images, there are intrinsic limitations to recognizing objects using solely visual information, especially in scenarios where the objects are occluded or partially visible. Indeed, RGB data are extremely biased toward the texture and color of the objects, encoding only a limited amount of geometry information. Moreover, it is sensitive to variations in lighting and other environmental conditions. For instance, if the lighting conditions change significantly it can cause a notable shift in the appearance of an object within an RGB image, making it more challenging for the model to accurately recognize the object. Using additional modalities, such as depth data or data from other sensors, can help to overcome these limitations and improve the accuracy and robustness of OR systems.

Challenges description. RGB-D (Kinect-style) cameras quickly became widespread in robotics vision due to their low cost and the wealth of visual data they provide. This sensor is composed of an RGB (red, green, blue) camera and a depth sensor, allowing it to capture both color and depth information about an object. The use of depth images provides complementary geometric information to the standard vision system. Moreover, it is the ideal workaround for the inherent limitations caused by the loss of data produced by projecting the three-dimensional world onto a two-dimensional image plane. In the field of RGB-D OR, the common pipeline involves two convolutional neural network (CNN) streams, operating on RGB and depth images respectively, as feature extractors.

However, the need for a large-scale dataset of depth images forced the vision community to find practical workarounds. Indeed, considerable effort has been spent on methods to colorize depth images in order to reuse the existing RGB pre-trained CNNs [42], while research into appropriate strategies to extract and combine features from the two modalities has been neglected. Indeed, several methods simply extract features from a specific layer of the two CNNs and combine them through a fully connected or a max pooling layer. We argue that these strategies are sub-optimal because (a) they assume that the selected layer always represents the best abstraction level to combine RGB and depth information and (b) they do not exploit the full range of information from the two modalities during the fusion process.

While the lack of an appropriate RGB-D fusion strategy may be a challenge for deploying RGB-D OR systems in robotics vision applications, it is not the main limiting factor. A larger concern is the amount of annotated data that is required to train CNNs for each new task, which can be costly and time-consuming, representing the main bottleneck. An appealing workaround that does not require manual annotation is to generate a large synthetic training set by rendering 3D object models with computer graphics software like Blender [4]. However, the difference between the synthetic (source) training data and the real (target) test data severely undermines the recognition performance of the network. This problem, known as the synthetic-to-real shift is an example of the broader domain shift issue introduced earlier in this section. This particular shift is well explored in the context of RGB data, by the research field called Unsupervised Domain Adaptation (UDA). This field has seen significant growth in recent years [43] and has produced numerous strategies for reducing the gap between source and target. However, these existing DA strategies often make the assumption that the data come from a single modality, which can lead to sub-optimal results when working with multi-modal data, as it ignores the natural relationships between different modalities.

Furthermore, it is important to recognize that multi-modal research is not limited to the use of RGB-D data, and can be easily extended to incorporate a wide range of innovative sensors. Indeed the event-based camera, or Dynamic Vision Sensor (DVS), presents several characteristics that make it ideal for the robotics context. This type of sensor belongs to the family of bio-inspired devices, where each pixel asynchronously emits an output, called *event* when it detects a local brightness change. This mechanism allows event-based cameras to operate at a very high dynamic range, high temporal resolution, and low latency, with minimal power consumption (more details in Section 2.3). As previously mentioned, the adoption of new sensors often comes with the challenge of limited data availability, hindering the full potential of deep learning algorithms for various tasks. This is particularly true for neuromorphic cameras, which are still relatively new and expensive, thus limiting the amount of available data. To overcome this issue, event camera simulators such as ESIM [44] provide an alternative solution by generating simulated event data. However, a crucial question remains unanswered: *how effectively does simulated data generalize to real-world scenarios?*

1.2 Recognize Actions in First Person videos

Since the infancy of computer vision, recognizing human actions from videos has been one of the most critical challenges. The ability to accurately and automatically recognize the actions performed by an individual or group of people has a significant impact on a wide variety of applications, including security, surveillance, and autonomous driving. Historically, the majority of research has focused on third-person action recognition, an area where significant progress has been made and where commercial products are now being released. In recent years, advances in wearable technology have sparked interest of the computer vision community in FPAR [10, 34, 45–53], both due to the challenges it presents and its potential for real-world egocentric vision applications, such as wearable sports cameras, human-robot interaction, and human assistance. This research area allows for a more direct study of human behavior, providing an opportunity to learn how humans perceive the world and interact with the environment.

Challenges description. Despite the wealth of visual information captured by wearable cameras, the transition from a third-person, fixed camera perspective to a first-person perspective introduces significant challenges. Specifically, the camera's motion during the recording of a first-person video, known as *egomotion*, can greatly alter the appearance of actions, making it difficult for the action recognition models to accurately classify them. Moreover, the presence of strong occlusion caused by the user's hand or arm during object manipulation, along with the existence of multiple and repetitive objects in the scene, further impedes the efficient localization and classification of the object of interest.

In addition to the aforementioned challenges, another significant limitation in this context is the presence of *environmental bias* [37, 54–57]. It refers to the tendency of RGB-based networks to heavily rely on the specific environment in which activities are recorded. Consequently, when actions are performed in unfamiliar or unseen surroundings, where the training and test data do not share the same distribution [58], RGB-based models struggle to recognize these actions effectively. This is mainly caused by the propensity of appearance-based networks to focus primarily on background cues and objects' texture, which are often uncorrelated with the action being performed and therefore vary significantly across different environments. As a result, appearance-free modalities, such as motion, have become the favored choice in current egocentric vision systems. However, the optical flow used in this setting is computed from RGB frames by solving expensive optimization problems (TV-L1 algorithm [59]), introducing significant test-time computations [60], making this modality totally impractical in online scenarios and preventing state-of-the-art performance from being achieved in real-world settings.

1.3 Contributions

The main contributions of this work are in the field of multi-modal object recognition and multi-modal egocentric action recognition. The specific contributions of this research are as follows:

- **a novel end-to-end architecture for RGB-D object recognition called recurrent convolutional fusion (RCFusion) [61],** it uses two streams of convolutional networks to extract RGB and depth features from multiple levels of abstraction. These features are then combined through a recurrent neural network (RNN), resulting in the generation of compact and highly discriminative multi-modal features.
- **a novel RGB-D DA method that reduces the synthetic-to-real domain shift [3]** by exploiting the inter-modal relation between the RGB and depth image. It consists of training a convolutional neural network to solve, in addition to the main recognition task, the pretext task of predicting the relative rotation between the RGB and depth image.
- **an alternative way of answering a very recent research problem regarding how to bridge Sim-to-Real gap arising from event generation [62].** We show that Unsupervised Domain Adaptation (UDA) techniques working at feature level are an effective way of tackling this issue, w.r.t. previous work that act on the input level. Moreover we propose a multi-view approach to deal with event representations, which outperforms existing methods and proved to work well in conjunction with other UDA strategies.
- **a single stream architecture for FPAR, called SparNet [63].** One of its key features is the integration of appearance and motion features through a set of self-supervised pretext tasks. These pretext tasks allow SparNet to estimate the motion information associated with a single static input image, enabling joint learning of appearance and motion features. This unique approach results in a lightweight architecture that can be trained in a single stage.
- **a new cross-modal loss called Relative Norm Alignment (RNA) loss [64].** We address the issue of differences in the marginal distributions of modalities that can hinder the training process and lead to suboptimal performance. We

introduce the RNA loss, which effectively tackles this problem by aligning feature norms across modalities in the Domain Generalization (DG) setting and across domains in the UDA setting. By ensuring balanced feature norms, the RNA loss enables models to better exploit the synergies and complementarities between different modalities, resulting in improved performance and adaptation capabilities.

- **the first event-based egocentric action recognition dataset, called N-EPIC-Kitchens [65].** We conducted a comprehensive comparative analysis to investigate the significance of motion information in the context of action recognition. To this end, we introduced and evaluated two novel approaches, namely E^2(GO) and E^2(GO)MO, specifically designed for event data. These approaches emphasize motion information and yielded competitive results compared to the computationally expensive optical flow modality.

1.4 Outline

In **Chapter 2**, we present a comprehensive survey of multi-modal learning, focusing on object recognition and first-person action recognition. In Section 2.2, we delve into the literature on Unsupervised Domain Adaptation (UDA) and Domain Generalization (DG) techniques and examine their applications in these tasks. Furthermore, in Section 2.3, we introduce a novel camera technology that has garnered significant attention in the robotics and computer vision communities due to its promising capabilities. Lastly, Section 2.4 provides an overview of the datasets that will be used for the experimental evaluation.

Chapter 3 describes the thesis contribution in the field of multi-modal object recognition. Section 3.1 presents **RCFusion**, a novel method for multi-modal fusion that exploits features from multiple hidden layers of CNNs for both modalities and fuses them using a recurrent neural network. In Section 3.2, we present **Relative Rotation**, a method for multi-modal unsupervised domain adaptation, based on self-supervised pretext task, and jointly we introduce a new synthetic dataset, called **synROD**. Lastly, in Section 3.3, we present **MV-DA4Event**, a solution that addresses the challenges arising from the simulated to real shift. We demonstrate the effectiveness of unsupervised domain adaptation in overcoming these challenges and highlight the significance of modalities in the context of multi-modal object recognition.

Chapter 4 showcases our contributions in the context of first-person action recognition. In Section 4.1, we present **SparNet**, a novel technique that enhances the ability of existing models to extract motion information. This is achieved by introducing an ad-hoc self-supervised pretext task. Next, in Section 4.2, we introduce **RNA**, a new loss function that effectively improves the overall generalization and adaptability of networks for fine-grained action recognition tasks. Furthermore, in Section 4.3, we propose **N-EPIC-Kitchens**, an extension of the existing first-person dataset [12] that includes event data. We also propose a strategy to adapt and reuse the existing model to leverage this new modality, providing a viable alternative to optical flow.

The thesis concludes with a summary discussion and remarks on possible future directions of research in **Chapter 5**.

1.5 Publications

Here is a list of publications categorized into the two main topics of OR and FPAR. The publications are further classified into journal and conference publications. Please note that a few of these published papers, marked with the symbol (†), are not included in this thesis.

Chronological list of Object Recognition Journal publications:

- Loghmani, M. R., **Planamente, M.**, Caputo, B., & Vincze, M.
  Recurrent convolutional fusion for RGB-D object recognition.
  IEEE Robotics and Automation Letters, 4(3), 2878-2885 – RAL 2019.

- Loghmani, M. R., Robbiano, L., **Planamente, M.**, Park, K., Caputo, B., & Vincze, M.
  Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition.
  IEEE Robotics and Automation Letters, 5(4), 6631-6638 – RAL 2020.

- **Planamente*, M.**, Plizzari*, C., Cannici*, M., Ciccone, M., Strada, F., Bottino, A., Matteucci, M., & Caputo, B.

Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation.
IEEE Robotics and Automation Letters, 6(4), 6616-6623 – RAL 2021.

Chronological list of Object Recognition Conference publications:

- Loghmani, M. R., **Planamente, M.**, Caputo, B., & Vincze, M.
  Recurrent convolutional fusion for RGB-D object recognition.
  IEEE/RSJ International Conference on Intelligent Robots and Systems – IROS 2019.

- Loghmani, M. R., Robbiano, L., **Planamente, M.**, Park, K., Caputo, B., & Vincze, M.
  Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition.
  IEEE/RSJ International Conference on Intelligent Robots and Systems – IROS 2020.

- **Planamente*, M.**, Plizzari*, C., Cannici*, M., Ciccone, M., Strada, F., Bottino, A., Matteucci, M., & Caputo, B.
  Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation.
  IEEE/RSJ International Conference on Intelligent Robots and Systems - IROS 2021.

- (†) Cannici*, M., Plizzari*, C., **Planamente*, M.**, Ciccone, M., Bottino, A., Caputo, B., & Matteucci, M.
  N-rod: A neuromorphic dataset for synthetic-to-real domain adaptation.
  In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1342-1347) – WCVPR 2021.

Chronological list of Egocentric Action Recognition Journal publications:

- (†) Goletto*, G., **Planamente*, M.**, Caputo, B., & Averta, G.
  Bringing Online Egocentric Action Recognition Into the Wild.
  IEEE Robotics and Automation Letters, 8(4), 2333-2340 – RAL 2023 (also presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems – IROS 2023).

The following works are currently under submission in a peer-reviewed journal:

- **Planamente, M.**, Plizzari, C., Peirone, A. S., Caputo, B., & Bottino, A.
  Relative Norm Alignment for Tackling Domain Shift in Deep Multi-modal Classification
- (†) Peirone*, S. A., Goletto*, G., **Planamente*, M.**, Bottino, A. Caputo, B., & Averta, G.
  Conceptual Activity-Centric Zones for Egocentric Action Recognition.

Chronological list of Egocentric Action Recognition Conference publications:

- (†) **Planamente, M.**, Russo, P., & Caputo, B.
  Leveraging over depth in egocentric activity recognition.
  First Italian Conference on Robotics and Intelligent Machines – I-RIM 2019. **(Best Paper Award finalists)**.
- **Planamente, M.**, Bottino, A., & Caputo, B.
  Self-supervised joint encoding of motion and appearance for first person action recognition.
  In 2020 IEEE 25th International Conference on Pattern Recognition (pp. 8751-8758) – ICPR 2021.
- **Planamente*, M.**, Plizzari*, C., Alberti, E., & Caputo, B. (2022).
  Domain generalization through audio-visual relative norm alignment in first person action recognition.
  In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1807-1818) – WACV 2022.
- (†) **Plananamente*, M.**, Plizzari*, C., & Caputo, B. (2022).
  Test-time adaptation for egocentric action recognition.
  In Image Analysis and Processing International Conference (pp. 206-218)– ICIAP 2022.
- Plizzari*, C., **Planamente*, M.**, Goletto, G., Cannici, M., Gusso, E., Matteucci, M., & Caputo, B. (2022).

E2 (go) motion: Motion augmented event stream for egocentric action recognition.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19935-19947) – CVPR 2022.

- (†) **Planamente, M.**, Goletto, G., Trivigno, G., Averta, G., & Caputo, B.
Toward Human-Robot Cooperation: Unsupervised Domain Adaptation for Egocentric Action Recognition.
In Human-Friendly Robotics 2022: 15th International Workshop on Human-Friendly Robotics (pp. 218-232). – HFR 2023
(Best Paper Award finalists).

- (†) Peirone, S. A., **Planamente, M.** Caputo, B., & Averta, G. (2023, May).
Test Time Adaptation for Egocentric Vision.
In Convegno Nazionale CINI sull'Intelligenza Artificiale – Ital-IA Workshops.

- (†) Neubert, J., **Plananamente, M.**, Plizzari, C., & Caputo, B.
LCMV: Lightweight Classification Module for Video Domain Adaptation
In Image Analysis and Processing International Conference – ICIAP 2023.

1.5.1 Technical Report

The following is a chronological list of technical reports related to the EPIC-Kitchens competition, which was presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021-2022). Our team achieved the third place in this competition for two consecutive years.

- (†) Plizzari*, C., **Planamente*, M.**, Alberti, E., & Caputo, B. (2021).
Polito-iit submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition.

- (†) **Planamente, M.**, Goletto, G., Trivigno, G., Averta, G., & Caputo, B. (2022).
PoliTO-IIT-CINI Submission to the EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition.

Chapter 2

Background and Related Work

The objective of this chapter is to offer a comprehensive survey of the literature on multi-modal learning, with a specific emphasis on object recognition and first-person action recognition tasks. Specifically, we delve into an overview of the commonly employed techniques that facilitate the practical implementation of these algorithms in real-world scenarios, highlighting the crucial role played by a model's ability to generalize and adapt. Moreover, this chapter provides an introduction to a novel camera technology known as event camera, which has recently attracted attention from the robotics and computer vision communities due to its promising capabilities

2.1 Learning to recognize: From Object to Action

Object recognition is a crucial aspect of robotics, as it enables robots to interact with their surroundings and perform various tasks. However, recognizing objects is just the starting point toward achieving more advanced robotic capabilities. To operate effectively in human environments, robots need to not only recognize objects but also understand human actions and intentions. This understanding allows robots to learn from humans and collaborate with them to efficiently and safely accomplish tasks. In this regard, the task of human action understanding plays a fundamental role. With the recent proliferation of wearable devices, the task of action recognition (EAR) has become particularly interesting and relevant for various computer vision and robotics applications. Wearable devices provide a unique perspective for capturing first-person visual data, enabling a more immersive and natural understanding of human actions. In this section, we will explore the literature related to object recognition, particularly from RGB-D data, and action recognition from both third and first-person perspectives. Additionally, we will provide an overview of the literature on event cameras, which offer distinct characteristics and have gained interest in the field of computer vision.

2.1.1 Robotics Vision: Object Recognition from RGB-D data

The literature on RGB-D object recognition can be broadly classified into three macro groups. The first group, referred to as the "traditional" group, primarily encompasses methods that emphasize the utilization of hand-crafted features. These methods involve the manual design and selection of specific characteristics to enhance the object recognition process, allowing for a detailed analysis of RGB-D data. The second group focuses on exploring methods that effectively leverage the multi-modal relationship between RGB and depth data. This group primarily consists of methods based on CNNs, which aim to exploit the complementary information provided by both RGB and depth modalities. Lastly, colorization techniques are employed in the third group to maximize the utilization of image-based pretraining for depth data. This group acknowledges the challenges posed by the absence of a robust pretraining model specifically tailored for depth data. The colorization techniques aim to bridge this gap by leveraging existing pretraining models primarily developed for RGB data, thereby enhancing the representation and understanding of depth data for improved object recognition.

Regarding the first group, Lai et al. [1] proposed a benchmark on a new RGB-D object recognition dataset, by combining spin image and SIFT. Browatzki et al. [66] introduce a 2D/3D dataset with a time of flight camera and then fused 2D features (SUFT, PHOG, Self-Similarity, and Color Histograms) and 3D features (SC3D, Depth Buffer, Shape-Index Histograms, and MD2 Shape Distributions) to predict the object classes. A set of kernel features to capture diverse yet complementary cues, resulting in a significant improvement in the performance of RGB-D object recognition, is introduced in [67]. However, these methods heavily rely on expert knowledge and lack fine generalization. Therefore, some researchers explore learning-based RGB-D object recognition methods. Blum et al. [68] designed a learning-based feature descriptor that integrates RGB and depth information into one consensus descriptor vector. Bo et al. [69] introduced hierarchical matching pursuit (HMP) to encode the feature representations of RGB-D inputs in an unsupervised manner. Asif et al. [70] calculated classification probabilities at pixel-level, surfel-level, and object-level from RGB-D data using the random forest method, and fused these probabilities to predict the classification results.

The second group primarily consists of approaches that repurpose deep learning architectures [71] trained on large RGB image databases, such as ImageNet. Specifically, certain studies [72, 71] have proposed the utilization of two-stream convolutional neural networks (CNNs) to learn features separately from RGB and depth modalities. Subsequently, the two embeddings are combined for the final prediction. Following these solutions, RGB and depth features are learned independently and connected only at the final stage through a fully connected layer classifier. However, while this approach is straightforward to implement, it fails to fully exploit the complementarity between different modalities. In contrast, Wang et al. [73] introduced a multi-modal layer into a CNN-based multi-modal learning network to discover discriminative features for each modality and exploit the complementary relationship between the two modalities. Additionally, Wang et al. [74] obtained multi-modal features by employing a pair of deep neural networks to separate the shareable and model-specific information from the extracted RGB and depth features. These methods combine the two modalities by processing features extracted from a single convolutional neural network and rely on a multi-stage optimization process to obtain individual and contextual information from both modalities. In contrast, more sophisticated approaches [73, 74] aim to enhance the network's capability to jointly learn and utilize the most discriminative features for each modality, while also leveraging the complementary relationship between the two. More recently, Zhang et al. [75] proposed a multi-modal fusion and projection (MMFP) module to address the issue of poor-quality depth images. The MMFP module reweights the contribution of each modality and combines the RGB and depth modalities using an attention mechanism, eliminating interference from irrelevant information while integrating the relevant information effectively. Differently [76] proposes ad-hoc regularization strategies for RGB-D data. These various approaches demonstrate ongoing efforts to enhance the integration of RGB and depth modalities for object recognition, leveraging new techniques to improve feature extraction and exploit the complementarity between different modalities.

The last macro group focuses on identifying optimal colorization strategies to effectively leverage the transfer of knowledge from RGB image pre-training to depth images. Various hand-crafted colorization techniques have been proposed to map the raw depth value of each pixel [72], derived physical quantities such as position and orientation [77], or local surface normals [69] to corresponding colors. In contrast, Carlucci et al. [78] introduced a learning-based approach to colorize depth images by training a colorization network. However, it is worth noting that some of these colorization schemes may be computationally expensive.

In addition to object recognition, the utilization of RGB-D data has been employed in various other contexts and tasks, such as scene recognition [79], object detection [80], pose estimation [81, 71, 82], activity recognition [83, 84], and hand gesture analysis [85, 86], requiring solutions that are sometimes distinct from those presented thus far.

2.1.2 Action Recognition

Action recognition (AR) is a type of computer vision task that involves analyzing videos and identifying the actions that are being performed in them. It is similar to image classification, in that it involves associating samples (in this case, video clips) with labels (in this case, actions). However, action recognition is a more complex task because it involves analyzing the motion and changes that occur over time in the video, rather than just the static content of an image.

AR approaches may be classified into four broad categories: 2D convolution-based [87–91, 47, 92, 93] and 3D convolution-based [94, 95, 49, 36, 96, 97, 87, 37, 98], transformer-based architectures and hybrid solutions that involve approaches that combine elements from different categories, such as using a combination of 2D and 3D convolutions or combining transformer architecture with convolutional layers.

The first group of AR approaches uses 2D convolutional neural networks (CNNs) to analyze the spatial features of the video frames. It is generally complemented with other modules enabling the possibility to learn temporal information. Recurrent neural networks, such as long short-term memory (LSTM) networks and their variations, are commonly used to capture temporal dependencies in video data [99, 34, 45, 46]. Differently, the authors of [91] proposed the Temporal Relation Network (TRN) a module designed to learn and reason about temporal dependencies between video frames at multiple time scales. Other approaches like Temporal Shift Module (TSM) [90] or its variant Gate-shift Module (GSM) [93, 100] use a temporal shift operator enabling a simple 2D convolution-based architecture to learn temporal features. The use of 3D convolutions was proposed as an alternative in [94, 96] to learn spatial and temporal relations simultaneously, even if they often introduce more parameters, requiring pre-training on large-scale video datasets [94]. The transformer architecture originally introduced for natural language processing tasks [101] is a model designed to capture long-range dependencies between words in a sentence by using a sequence of self-attention blocks. In recent years, researchers have attempted to extend the transformer architecture to the vision and video domains by adapting it to work with images and videos, respectively [102–107]. To do this, they have proposed various approaches such as tokenizing non-overlapping patches of the image and using data augmentations and a student-teacher training scheme to improve data efficiency. In the video domain, various methods have been proposed to incorporate temporal information, such as TimeSformer [108], ViViT [109], and VideoSwin [110], which use temporal attention schemes to compare patches in different frames. Another approach, MViT [111], aims to reduce the computational cost of video transformers by using a local pooling operation to progressively reduce the number of tokens while increasing the channel dimension.

The last group of AR approaches consists of hybrid solutions, which are approaches that combine elements from the first three groups. For example, in Girdhar et al.'s work [112], the authors propose an architecture that combines 3D CNNs with a self-attention mechanism for final prediction. In contrast, Tai et al. [113] integrate transformer architecture with recurrent mechanisms in a unified model.

Another area of research in AR focuses on the efficiency of action recognition models, aiming to find architectures that can achieve good performance while using fewer parameters. This can be achieved through the use of techniques such as Neural Architecture Search (NAS), as demonstrated by the work of Kondratyuk et al. [114], who present Mobile Video Networks (MoViNets), a family of computation- and memory-efficient video networks that are optimized using NAS and are capable of operating on streaming video in real-time. There are also other previous approaches such as [115–120], which aim to find more efficient architectures that can perform well with fewer parameters. Although all these architectures aim at implicitly modeling motion, most of them still mix video frames with the externally estimated optical flow. While this improves the overall performance, it also requires precomputing the flow, making these approaches impracticable in online settings. In addition, two-stream approaches come at the cost of increased model complexity and the number of parameters.

2.1.3 First Person Action Recognition

In the field of FPAR, many of the networks used are inherited from the third-person action recognition field [94, 90, 88, 121]. However, the unique challenges of the egocentric point of view have led researchers to investigate solutions specifically tailored to this task. Several approaches make use of object or hand detection mechanisms to guide the action recognition task [122, 122–124, 89, 125–128]. However, these methods require specific annotation of the hand or object and can introduce latency during the inference process. Recently, Shan et al. [129] developed a hand-object detector to locate the active object. However, this solution is also not lightweight and fast, limiting its use in online scenarios. Alternative approaches that have been gaining traction include the use of attention mechanisms guided by elements such as gaze information, active objects, and/or the user's hands [10, 130, 45, 34, 131–133]. These mechanisms have been shown to greatly enhance the performance of FPAR systems.

In addition to this, multi-modal analysis plays a crucial role in the field of FPAR. The use of a two-stream solution, in which visual information is combined with optical flow, is a common practice in many studies [37, 45, 47, 46]. Additionally, a particular focus on the utilization of alternative modalities to improve the robustness and accuracy of predictions. One such alternative modality that has been explored in this context is audio signals. Studies such as Kazakos et al. [47] have demonstrated the effectiveness of incorporating audio information in action recognition tasks, owing to the rich sound information present in hand-object interactions and the proximity of sensors to the sound source. Other modalities such as Electromyography (EGM) signal, Inertial Measurement Unit (IMU), depth, and tactile data are also being explored in this field [134–139].

More generally, the egocentric vision has introduced a range of new challenges for the computer vision community, including human-object interaction [140, 141], action anticipation [142, 46, 143, 130], action recognition [47], and video summarization [144–146]. With the availability of large-scale datasets [12, 13], new tasks have been proposed, such as wearer's pose estimation [147] and egocentric video anonymization [148]. This trend is expected to continue with the recent release of Ego4D [139], a massive egocentric video dataset featuring over 3,000 hours of daily-life activity videos with accompanying audio, 3D meshes of the environment, eye gaze, stereo, and multi-view videos.

2.2 Learning to see across Domains and Modalities

The standard assumption in machine learning is that the distribution of training data is representative of the distribution encountered during testing. However, in practical scenarios, this assumption is often challenging to fulfill, leading to a phenomenon known as domain shift [58]. Domain shift refers to the situation where the underlying data distributions in the training and testing phases differ significantly, resulting in a discrepancy between their statistical properties. Maintaining the assumption of matching distributions between training and testing data is difficult due to various factors. For instance, in computer vision, changes in lighting conditions, viewpoints, weather, or the presence of occlusions can cause a shift in the data distribution. Other examples include variations in data collection setups, such as different sensors or modalities, which introduce discrepancies between training and testing domains. These discrepancies can lead to performance degradation when models trained on one domain are applied to another.While this problem has been extensively investigated in the context of image classification, its relevance transcends visual data alone. The challenge of domain shift extends to various data modalities, encompassing audio, depth, text, and even temporal domains. Adapting models to operate effectively across different modalities and diverse data domains is crucial for achieving generalization and robustness in various application domains. In the subsequent sections, we will first provide a formal definition of this problem, considering its implications across different domains and modalities. We will then offer a comprehensive overview of the existing literature, highlighting the various approaches and techniques proposed to mitigate the challenges posed by domain shift.

2.2.1 Problem Formulation

Domain adaptation is a subfield of machine learning that addresses the challenge of learning from a source domain and applying the acquired knowledge to a different target domain. The main objective is to leverage the knowledge gained from the source domain in order to improve the performance of a model when faced with data from different distribution. More specifically, a domain can be defined as a tuple $\mathcal{D} = \{\chi, \gamma, p(x \mid y)\}$, where χ represents the input or feature space, γ is the label space, and $p(x|y)$ represents the joint probability distribution over the input-label space pair $\chi \times \gamma$. The source domain S is composed by input-label pairs $\{(x_i^s, y_i^s)\}_{i=1}^n$, where n is the number of samples in the source. Similarly, the target domain T is composed by $\{(x_i^t, y_i^t)\}_{i=1}^m$, where m is the total number of samples. Notably, in unsupervised domain adaptation, the labels y_i^t are not available during training.

The primary objective of unsupervised domain adaptation is to reduce the distribution discrepancy between the source and target domains. This discrepancy refers to the differences in data distributions between the two domains. By reducing this discrepancy, we aim to enable the training of a model that performs well on the target samples. In essence, domain adaptation seeks to learn a generalized classifier that can effectively classify samples from the target domain, despite the presence of a shift between the distributions of the source and target domains. One key assumption in this context is that the label space remains consistent between the source and target distributions. This assumption implies that the same set of labels or classes exists in both domains, facilitating the transfer of knowledge from the source to the target domain. By leveraging this shared label space, unsupervised domain adaptation techniques aim to bridge the gap between domains and enable effective generalization to unseen target samples.

Domain Generalization (DG), is another research area with a similar objective to domain adaptation. However, in this setting, the target data is not available during training. DG leverages the availability of multiple source domains for training. More formally, during training, the model has access to one or more fully labeled source datasets $\mathcal{S}_1, \ldots, \mathcal{S}_m$, but no information is available about the target domain $\mathcal{T}$. The goal of domain generalization is to develop models that can generalize well to unseen target domains by learning robust representations that capture the common knowledge across the source domains while being invariant to domain-specific variations.

Multi-Modal Adaptation. Our objective is to explore the impact of leveraging multi-modal signals from both source and target data on the ability to bridge the gap between different domains. In our setting, each input x_i^d, where d denotes the domain (either source or target) and i represents the i-th sample in the dataset, consists of multiple modalities. Specifically, x_i^d can be expressed as a collection of modalities: $x_i^d = x_{i,m_1}^d, ..., x_{i,m_p}^d$, where p denotes the number of modalities present in the input. To handle these multi-modal inputs effectively, we employ separate feature extractors, denoted as $F_1, ..., F_p$, where each extractor corresponds to a specific modality. The multi-modal approach allows us to leverage the strengths of different modalities and promote cross-modal knowledge transfer, ultimately facilitating the adaptation and generalization across domains.

2.2.2 Unsupervised Domain Adaptation

UDA is a thoroughly investigated research domain in the field of image classification. The existing literature on UDA encompasses a diverse set of methodologies, which can be categorized into five primary groups for better organization and understanding. The first group consists of discrepancy-based methods [149–151] that aim to quantify the discrepancy between source and target data in the feature space. One widely employed metric in this group is Maximum Mean Discrepancy (MMD), which is minimized during network training to address the domain shift. Some approaches extend the use of MMD to multiple layers of the network [152], facilitating a comprehensive alignment of features across different domains. Researchers have proposed various variations of the MMD metric to enhance its effectiveness. For instance, Long et al. [149] introduced a multiple-kernel variant of MMD, while [153] suggested Central Moment Discrepancy (CMD) that emphasizes aligning the central moments of the first k orders, thereby further improving domain adaptation performance.

The second group of methods focuses on adversarial learning [154–159]. These methods employ an adversarial training framework to encourage the network to generate representations that are indistinguishable between the source and target domains, thus achieving domain-invariant representations. One popular technique used in adversarial-based unsupervised domain adaptation is the Gradient Reversal Layer (GRL). It involves adding a second discriminator head to the standard network architecture, which is trained to distinguish between source and target features. During training, the gradient of the discriminator branch is reversed, meaning that the feature extractor is encouraged to produce domain-invariant representations that cannot be easily distinguished by the discriminator. Furthermore, variations of this approach have been proposed, as discussed in [160], where multiple domain discriminators, one for each class, are utilized to align the conditional distributions. In the case of unlabeled target data, the probability of the classifier for the k-th class is used to weight the target features used as input for the k-th domain discriminator. Similarly, the metric discrepancy approach is also utilized at various levels, as explored in [161]. The second category of methods also includes the generative approach, which involves translating the source images into the style of the target domain. Several studies [162–166] fall under this category. In the initial works [162, 166], a Generative Adversarial Network (GAN) architecture was employed. The generator was trained to reconstruct the original source image, while the discriminator was trained to differentiate between the reconstructed source image and the target images. By maximizing the discriminator's loss with respect to the generator, the GAN produced images that closely resembled the style of the target domain, while still preserving the task-specific information from the source domain. To address the issue of features collapsing into uni-modal solutions, cycle consistency constraints were introduced [163, 164]. These constraints ensure that the generated source images, which have been transformed to match the style of the target domain, are translated back to the original style of the source domain.

The third group includes methods that show how to properly reuse the batch normalization layers to normalize source and target statistics [167–169]. One approach, called AdaBN [167], involves updating the population statistics of all BN layers with the statistics of the target domain as a post-processing step after training. In some studies [170–173], separate BN layers have been employed for the source and target distributions to align the domains during training. A variation of this approach is presented in [172], where the authors argue that deeper layers learn more abstract and ideally domain-invariant concepts, and thus should share BN layers. However, the lower-level layers, which are domain-specific, should still have separate BN layers. Their method, AutoDIAL, utilizes a mixing parameter to control the extent to which the distribution statistics of each BN layer should be shared across domains. This mixing parameter is learned as the training progresses. When the distribution statistics are completely shared, AutoDIAL functions as a single BN layer, whereas if the statistics are not shared, AutoDIAL acts as two separate BN layers.

The next group of methods focuses on leveraging self-supervised learning techniques to mitigate the domain shift [174–176]. More precisely, a network is trained to solve an auxiliary self-supervised task on the target (and source) data, in addition to the main task, to learn robust cross-domain representations. One commonly used self-supervised task is predicting the transformations applied to input images. Studies such as [176, 175, 177] employ tasks like jigsaw puzzle solving, rotation prediction, and patch location estimation in conjunction with the main task to facilitate learning. Furthermore, the recent success of the contrastive learning approach in representation learning has led the computer vision community to explore its application in unsupervised domain adaptation [178–180]. Contrastive learning aims to learn an embedding space where positive pairs are pulled closer together, while negative pairs are pushed apart. However, one challenge in applying this approach to unsupervised domain adaptation is finding positive samples for the target domain, considering that the target data are unlabeled. In [179], the authors propose using clustering on the target data and selecting samples that are closer to the centroids as positive pairs.

The final group of methodologies is dedicated to reducing classifier uncertainty by minimizing entropy. Within this category, several adaptation methods have been proposed [172, 181–184]. For instance, Information Maximization (IM) [185] combines entropy minimization with the maximization of source mutual information between the classifier outputs and feature representations. Another notable approach is Minimum Class Confusion (MCC) [186], which introduces a novel loss function that effectively reduces classifier uncertainties, particularly in scenarios involving pairwise class confusion. This improvement leads to substantial gains in transfer performance.

The approaches mentioned above, originally developed for standard image classification tasks, have also been extended to various video-related applications. These applications include action detection [187], segmentation [188], and classification [189, 37, 190–193]. Specifically, concerning unsupervised domain adaptation (UDA) for action recognition, a multi-level adversarial framework named TA^{3}N is proposed in [189]. TA^{3}N incorporates temporal relation and attention mechanisms to align the temporal dynamics of the video feature space. TCoN [192] employs a cross-domain co-attention mechanism to align feature distributions between the source and target domains. Additionally, [190] presents an approach that trains the network on an auxiliary self-supervised task. CoMix [57] introduces a contrastive learning framework for discriminative feature representation, while SAVA [194] addresses domain adaptation through clip order prediction as an auxiliary task. It is worth noting that these proposed methods primarily aim to extend existing image-based UDA solutions to the temporal dimension, often incorporating a combination of different approaches.

2.2.3 Domain Generalization

In the field of DG, the focus is on building models that can perform well on target domains without the availability of target domain data. This is achieved by utilizing knowledge from one or multiple source domains. The majority of the existing literature in this area, as for the UDA, focuses on image data methods [176, 195–200] and can be broadly classified into three main groups: aligning source domain distributions for domain-invariant representation learning [196, 198, 176], exposing the model to domain shift during training via meta-learning [201, 202], and augmenting data with domain synthesis [203].

Several works have explored the task of DG in the context of videos as well. For instance, in [204], a simple episodic training strategy is proposed to mimic the train-test domain shift during training. This strategy enhances the model's robustness to novel domains. In contrast, the authors of [196] extend adversarial autoencoders by incorporating Maximum Mean Discrepancy (MMD) measures. These measures are used to align the distributions among different video domains, while adversarial feature learning matches the aligned distribution to an arbitrary prior distribution. In [205], a causal approach is introduced to generalize between video domains. The method achieves this by training the model to recognize the same action performed in different backgrounds, thereby promoting domain generalization. More recently, the work presented in [206] introduces VideoDG, an Adversarial Pyramid Network specifically designed for video generalization. VideoDG captures local-relation, global-relation, and cross-relation features progressively, enabling effective generalization of video features. Furthermore, in the context of domain generalization using egocentric data, it's important to highlight a recent development – the creation of the Action Recognition Generalisation Over scenarios and locations dataset, or ARGO1M [207]. This dataset contains an extensive 1.1 million video clips gathered from the comprehensive Ego4D dataset, covering 10 scenarios and 13 locations. Research findings have demonstrated that recognition models face significant challenges when attempting to generalize across the 10 proposed test splits, each presenting an unseen scenario in a completely new location. This underscores the continuously evolving landscape and complexities within domain generalization involving egocentric data.

2.2.4 Multi-Modal Adaptation

The previously discussed methods have predominantly focused on single-modal data. However, the combination of multiple modalities in multi-modal data introduces new challenges. The nature of domain shift may affect different modalities in distinct ways, particularly when they are heterogeneous. Additionally, multi-modal approaches often take into account the temporal dimension alongside the various modalities. Therefore, there is a need to explore techniques that can effectively adapt to multi-modal data.

Despite the wide range of techniques for multi-modal unsupervised domain adaptation (UDA), there are common underlying principles. Firstly, many studies adapt existing UDA techniques developed for single-modal images to handle multi-modal scenarios [208, 209]. This approach leverages the knowledge and insights gained from UDA research to achieve similar results in multi-modal applications. Adversarial-based approaches, such as MDANN [210] and AUDA [211], focus on learning discriminative and domain adaptive features using an adversarial objective. These approaches have demonstrated effectiveness in cross-domain emotion recognition using audio-visual data and cross-media retrieval using images and text from different domains.

Secondly, these approaches aim to leverage the multi-modal nature of the data to enhance the inter-correlation between different modalities and achieve a more robust data representation. The objective is to construct a shared representation space that captures significant inter-modal relationships, facilitating seamless adaptation to the target domain without relying on labeled data. Often, this second group combines the proposed solutions with techniques from the first group, as referenced in [37, 212, 213]. For instance, in MM-SADA [37], they extend adversarial alignment to a self-supervised task based on modality correspondence. Co-training methods, such as DLMM [214] and XM-UDA [215], exploit the diverse properties of different modalities by treating the classifiers of various modalities as a set of teacher/student models trained with a curriculum learning approach. These methods have been applied to tasks such as event recognition using audio-visual data, fatigue detection using EEG signals and facial keypoints, and action recognition using RGB images and optical flow. Contrastive learning-based methods, such as STCDA [55] and the approach described in [56], utilize the complementarity of different modalities to regularize both cross-modal and cross-domain feature representations. They treat each modality as a view and perform contrastive learning across modalities and domains to align representations between source and target domains in each modality. CIA [212] employs cross-modal interaction and generative modeling to align cross-domain representations. Despite the ongoing research in this field, the methods often become highly specific to particular tasks and/or modalities, lacking a true comparison of a method designed for universal modalities (not only visual).

2.3 Learning to see from Event data

Event-based cameras, known as Dynamic Vision Sensors (DVS) [216–220], employ a bio-inspired sensing approach that closely mimics the operations of biological retinas. Unlike conventional cameras, which capture complete images at a fixed rate determined by an external clock (e.g., 30 frames per second), event cameras detect asynchronous and independent brightness changes for each pixel in the scene. These cameras exclusively generate data when there is a change in brightness. Consequently, the output of an event camera consists of a variable data sequence of digital "events" or "spikes." Each event represents a predefined magnitude change in brightness (log intensity) occurring at a specific time t and pixel location (x, y). Mathematically, an event tuple can be represented as $E = (x, y, t, p)$, where x, y, t, and p denote the pixel location, time, and polarity (1-bit) indicating a brightness increase ("ON") or decrease ("OFF"). Each pixel stores the log intensity each time it emits an event and continuously monitors for a significant change from this stored value. When the change exceeds a threshold, the camera identifies an event. Event cameras are data-driven sensors, meaning that their output depends on the amount of motion or brightness change in the scene. As motion increases, more events are generated per second, as each pixel adjusts its sampling rate based on the rate of change in the monitored log intensity signal. These events are precisely timestamped with microsecond resolution and transmitted with sub-millisecond latency, enabling these sensors to respond quickly to visual stimuli.

Furthermore, event cameras offer several distinct advantages over standard cameras. Firstly, they provide high temporal resolution by swiftly monitoring brightness changes through analog circuitry. The read-out of events is digital, employing a 1 MHz clock, which enables the detection and timestamping of events with microsecond precision. This high temporal resolution allows event cameras to capture fast motions without suffering from motion blur, a common issue with frame-based cameras. Moreover, event cameras exhibit low latency as each pixel operates independently, eliminating the need for a global exposure time for the entire frame. As soon as a change in brightness is detected, it is immediately transmitted, resulting in minimal latency. In laboratory settings, event cameras have demonstrated latency of around 10 µs, and in real-world applications, the latency is typically sub-millisecond. Additionally, event cameras excel in terms of power efficiency. They transmit only the relevant brightness changes, eliminating the transmission of redundant data. This significantly reduces power consumption, as energy is only used to process pixels that undergo changes. At the die level, most event cameras consume around 10 mW of power, and there are even prototypes that achieve power consumption below 10 µW. Embedded event-camera systems, where the sensor is directly connected to a processor, have exhibited system-level power consumption (including sensing and processing) of 100 mW or less [31, 221, 222]. Moreover, event cameras offer a high dynamic range (HDR) that surpasses that of high-quality frame-based cameras. With a dynamic range exceeding 120 dB, event cameras can capture information across a wide range of lighting conditions, from moonlight to daylight. This exceptional dynamic range is achieved because the photoreceptors of event camera pixels operate on a logarithmic scale, and each pixel operates independently without the need for a global shutter. Similar to biological retinas, event camera pixels can adapt to both very dark and very bright stimuli, providing superior HDR capabilities.

These unique characteristics have garnered significant attention, initially within the robotics community and subsequently in the field of computer vision. The potential applications of event cameras are extensive, spanning various domains such as real-time interaction systems, robotics, and wearable electronics [223], where operating under uncontrolled lighting conditions, minimizing latency, and optimizing power consumption are critical considerations [224]. Event cameras offer distinct advantages over other sensing modalities in numerous scenarios. They excel in object tracking [225, 226], surveillance and monitoring [227], and object/gesture recognition [228, 229, 31]. Additionally, they prove beneficial for depth estimation [230, 231], structured light 3D scanning [232], optical flow estimation [233, 234], HDR image reconstruction [235, 236], and Simultaneous Localization and Mapping (SLAM) [237–239]. The field of event-based vision is continuously expanding, and as event cameras become more widely accessible, we can anticipate the emergence of new applications, such as image deblurring [240][28] and star tracking [241, 242] [243].

Events captured by neuromorphic cameras possess limited information when examined in isolation. They primarily signal changes in brightness at specific spatio-temporal coordinates. To derive meaningful predictions, computer vision algorithms necessitate an aggregation mechanism that correlates events within a neighborhood in space and time. Consequently, the research community has identified two prevailing paradigms for event processing, which diverge based on their approaches to event consumption during computation.

The first paradigm, known as *event-by-event computation*, involves processing each event as it occurs. This incremental and asynchronous approach enables the algorithm to update its output in real-time, achieving minimal reaction times. Such algorithms often rely on an internal state that is continuously updated upon the arrival of each event. This state represents the algorithm's understanding of the scene's content and its temporal evolution. Spiking Neural Networks (SNNs) [244] are the leading approach for neural systems that perform asynchronous spike-based computation. These networks consist of individual neuron-like units that respond asynchronously to incoming spike events. They accumulate spikes from other neurons and fire when a sufficient amount of relevant information is gathered. The SNN's internal memory is represented by the neurons' membrane potential, which is updated event-by-event as each event arrives asynchronously. SNNs have been applied in various event-based processing tasks, such as edge detection [245, 246] and object and hand-gesture recognition [247–249]. Due to their biological inspiration, SNNs are typically trained using unsupervised, biologically inspired learning rules [250, 251]. However, recent studies [248, 252–254] demonstrate that adopting hybrid approaches that combine traditional gradient-based learning methods can lead to superior performance in SNNs.

Filtering algorithms represent another prevalent category of approaches within the event-by-event paradigm. They are specifically designed to handle incomplete and potentially noisy sets of observations, incorporating the concept of a continuously updated state as new observations are received. These algorithms find applications in various fields, including Simultaneous Localization and Mapping (SLAM) algorithms [236, 255, 256], widely used in conventional computer vision approaches. They are also utilized in noise filtering mechanisms [257, 258] and image and video reconstruction algorithms [259, 260], particularly in converting event camera output to grayscale. Deterministic filters have been employed for event-based artificial neural networks [252, 261, 221, 262] to implement asynchronous convolution and feature extraction [263, 264]. These filters take advantage of the sparse encoding capabilities of event cameras, enabling fast operations with minimal computational cost. Instead of processing the entire image, they focus on the local neighborhoods surrounding incoming events.

The batch-based paradigm is the second approach, where algorithms wait for a batch of events to arrive and process them all together, prioritizing performance over response time. Offline processing of the event stream often involves a sliding window approach, where the batch is constructed based on either a fixed number of events or a specific time period for each window. Similar to the event-by-event approach, batch-based methods may employ an internal state to extend the context beyond the batch and integrate information from previous computations. In batch-based approaches, the input stream is typically converted into structured representations that provide richer information by enabling the correlation of events in space and time. While some pre-processing mechanisms retain the asynchronous and sparse encoding [265–268], others convert the event stream into densified representations

[269–271]. The concept of time surfaces [263, 264] is employed by several batch-based algorithms to extract dense representations that capture local spatio-temporal motion footprints from each event. These time surfaces find diverse applications, including 3D stereo reconstruction [272, 273], stereo depth estimation [274], SLAM [265], corner detection and tracking [275–277], as well as object classification [263, 264].

Some batch-based methods take a different approach by avoiding the creation of dense intermediate representations and instead leveraging the sparse encoding capabilities of event cameras. These methods treat events as vertices in a graph that is interconnected based on local spatio-temporal neighborhoods. Graph-based techniques have recently been utilized in motion segmentation algorithms [265], and graph neural networks have demonstrated success in tasks like object classification [266, 267] and action recognition [267, 268]. Another similar approach treats event streams as 3D point clouds, where the temporal dimension replaces the depth dimension. Deep learning networks such as PointNet [278] and PointNet++ [278] have shown promising results when applied to small temporal windows of events in tasks like object recognition, semantic segmentation [279], and gesture recognition [280].

The most popular methods involve converting the event stream into dense representations known as event frames. These representations resemble conventional frames and can be easily integrated into conventional computer vision pipelines. Deep learning approaches based on these grid-like encodings have been applied to various tasks, including object classification [281, 269] and detection [282], semantic segmentation [283], depth and optical flow estimation [234, 284, 285], as well as image reconstruction [286–288]. As this thesis places a particular emphasis on grid-like event representations, the following section will provide an in-depth exploration of this area. Detailed information and insights will be presented to offer a comprehensive understanding of grid-based event representations.

2.3.1 Grid-Like Event Representations

Given a stream of asynchronous events $\mathcal{E} = e_i = (x_i, y_i, t_i, p_i)i = 1^N$, the process of extracting a grid-like representation can be described as converting $\mathcal{E}$ into a volume $\mathcal{R}_{\mathcal{E}} \in \mathbb{R}^{H \times W \times F}$ with F features. Various representations have been proposed for grid-like event representations, offering different approaches to computing and aggregating pixel features. Traditionally, grid-like representations are hand-crafted, meaning the transformation mapping the event stream to $\mathcal{R}_{\mathcal{E}}$ is fixed and independent of the task. However, recent works [269, 271] propose training them with the rest of the network to learn task-specific representations.

Within the hand-crafted representation group, several methods have been developed for grid-like event representations. The event counts representation [234, 289] focuses on aggregating event sequences based on their cardinality, without considering temporal information. On the other hand, the Surface of Active Events (SAE) [290, 233] method preserves recent temporal information by tracking the timestamp of the last event received at each pixel. An extension of SAE, known as time surfaces, incorporates time-modulated kernel functions into the SAE representation. In the Brightness Increment Image (BII) [291], polarity information is utilized, where the pixel feature is computed by summing the polarity values of all events that occurred at the same location. Another recent extension of SAE, proposed by Kim et al. [292], exploits the spatial neighborhood around each pixel to suppress noisy event contributions and quantize temporal features, effectively removing the time scale from temporal delays. In [264] the authors propose the Histograms of Time Surfaces (HATS), a two-channel representation format. To construct HATS, the initial event stream grid is divided into C cells, each with dimensions $K \times K$ pixels. For each polarity p and each cell c, a grid of $(2\rho+1) \times (2\rho+1)$ time surface histograms $\mathbf{h}_{c,p}$ is computed based on the events generated by the pixels within the cell. These histograms are then normalized and rearranged according to their originating cell position, resulting in two channels—one for each polarity. It is important to note that HATS typically sacrifice temporal resolution, as the entire temporal window is condensed into a single frame, thereby losing fine-grained temporal information. Voxel grid representations [8] have gained substantial popularity in deep learning applications. They serve as inputs to deep architectures for various tasks, including optical flow, depth estimation, egomotion prediction [234, 284], object detection [282], classification [269], and image reconstruction [287, 288, 293]. This representation, also referred to as event volume, discretizes the time domain into a traditional image format. The voxel grid consists of a fixed number of channels, denoted by B (where $F = B$), and events from the event stream are inserted into the grid using temporal interpolation. The resulting voxel grid can be mathematically expressed as:

$$\mathcal{R}^{\text{vox}}\mathcal{E}(x,y,b) = \sum i = 1^N p_i k_b(x - x_i) k_b(y - y_i) k_b(b - t_i^*), \tag{2.1}$$

where b represents the channel number, t_i^* denotes the timestamps scaled into the range $[0, B-1]$, and $k_b(a) = \max(0, 1 - |a|)$. zhu2019unsupervised In the second group, which pertains to learnable representations, we find the Event Spike Tensor (EST) [269]. This representation is designed to be end-to-end trainable and operates similarly to a voxel grid. However, the key distinction lies in the use of a learned kernel function $\mathcal{K}$ to compute the contribution of each event. The kernel function is determined by a multi-layer perceptron (MLP) network. The computation of the final representation is given by:

$$\mathcal{R}^{\text{EST}}\mathcal{E}(x,y,b,p) = \sum e_i \in \mathcal{E}^{(x,y,p)} \hat{t}_i \cdot \mathcal{K}\left(\hat{t}_i - \frac{b}{B-1}\right), \tag{2.2}$$

Here, $\mathcal{E}^{(x,y,p)}$ represents the set of all events with polarity p received at the specific pixel (x,y), and $\hat{t}_i = \frac{t_i}{t_N}$ denotes the normalized event timestamp. A more recent contribution in this category is the MatrixLSTM representation proposed by the authors of [281]. MatrixLSTM shares similarities with EST but introduces a novel approach to compute task-dependent event surfaces. It employs a matrix of LSTM cells with shared parameters. Each cell processes the temporal sequence of events generated by a pixel (x,y), and the last output of the LSTM is utilized as the pixel feature. Optionally, the time window can be divided into bins to increase the number of output channels.

2.3.2 Data Scarcity Challenge in Event-Based Vision

In recent years, novel learning approaches based on standard computer vision algorithms operating on event data have emerged as competitive alternatives to traditional methods [289, 269]. However, training off-the-shelf deep learning algorithms typically relies on large amounts of data, which is still limited due to the novelty and high cost of neuromorphic cameras. To overcome this challenge, researchers have explored the use of event camera simulators [44], which generate reliable simulated event data. Nevertheless, a critical research question arises from this approach: How well do simulated data generalize to real data? Recent efforts have partially addressed this question in [294] and [295], where authors proposed techniques to reduce the "sim-to-real gap" by manipulating simulator parameters at the input level during the data simulation phase. These endeavors highlight the potential of simulated data to mitigate the data scarcity challenge. However, the generalizability of simulated event data remains an open research question that warrants further investigation.

The scarcity of data is particularly limiting in exploring new tasks where event data could offer significant benefits, both in terms of the information they encode and the unique characteristics of the event camera devices. One such context where event data holds immense potential is wearable computing, where the low-power, high temporal resolution, and asynchronous nature of event cameras align well with the requirements of wearable devices. However, collecting a comprehensive dataset for such new tasks can be challenging and time-consuming.

Therefore, addressing the lack of data is crucial not only for training deep learning models but also for exploring new applications where event data can have a significant impact. This limitation motivates the development of innovative techniques, such as data augmentation methods [296–299], transfer learning strategies [294, 292], and domain adaptation approaches [300, 301, 62], to effectively utilize limited event data and unlock the full potential of event-based vision in various domains, including wearable computing.

2.4 Dataset

We introduce two distinct sets of datasets: one dedicated to object recognition within the field of robotics, and another specifically designed for action recognition tasks, with a strong emphasis on egocentric vision. Although these groups vary in the modalities they encompass, they both offer valuable resources for validating and evaluating our multi-modal solutions. In the following sections, we offer comprehensive descriptions of each dataset, emphasizing the included modalities, annotations, and the specific tasks they cater to.

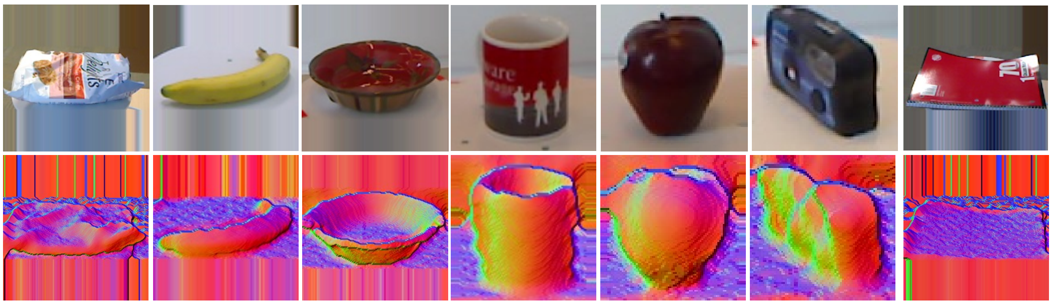


Fig. 2.1 Samples from the RGB-D Object dataset [1], showcasing two modalities: RGB (top) and depth images (bottom). The data has been processed according to the procedure outlined in [5]. The depth information is represented using surface normals, employing the best non-learned colorization method currently available.

2.4.1 Robotics Vision Datasets

We provide a detailed description of the robotic vision datasets employed in our validation process. These datasets serve as valuable resources for training, testing, and benchmarking the performance of algorithms in real-world robotic scenarios. By utilizing these datasets, we can quantitatively evaluate the performance of our proposed approaches in terms of accuracy and robustness. Additionally, the use of standardized datasets allows for fair comparisons with existing state-of-the-art methods in the field. This ensures that our evaluation results are reliable, reproducible, and relevant to the broader research community.

ROD. The RGB-D Object dataset [1] consists of 41,877 images representing 300 objects from 51 categories, including fruit, vegetables, tools, and containers as shown in Figure 2.1. Each object is recorded on a turn-table with the RGB-D camera placed at an approximately one-meter distance at 30°, 45°, and 60° angle above the horizon. This dataset has become a widely used benchmark for analyzing methods of RGB-D object recognition. For the evaluation, we use the standard experimental protocol defined in [67], where ten training/test splits are designed such that one object instance per class is excluded from the training set. The reported results show the average accuracy across all splits.

JHUIT-50. The RGB-D image dataset [6] used in this study comprises 14,698 images of 50 common workshop tools, including clamps and screw drivers, as shown

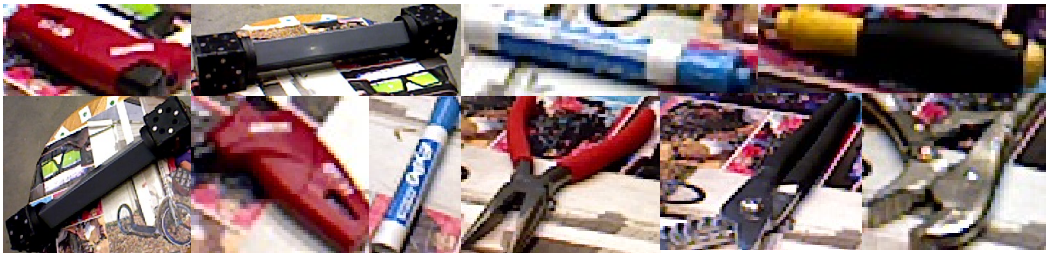


Fig. 2.2 Samples from the RGB-D image dataset [6] used in this study. The dataset comprises images of workshop tools, including clamps and screwdrivers.

in Figure 2.2. Despite its small number of object classes, this dataset presents a challenging task for instance recognition, as the objects are highly similar to each other. For the evaluation, we adopt the experimental protocol defined in [6], where the training data are collected from fixed viewing angles, while the test data are collected by freely moving the camera around the object.

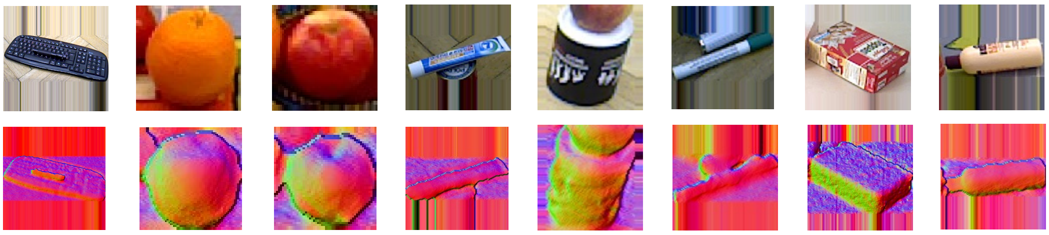


Fig. 2.3 Samples from the Object Clutter Indoor (OCID) dataset, showcasing two modalities: RGB (top) and depth images (bottom). The data has been processed according to the procedure outlined in [5]. The depth information is represented using surface normals, employing the best non-learned colorization method currently available.

OCID. The Object Clutter Indoor (OCID) dataset consists of 96 cluttered scenes that depict a variety of common objects, organized into three subsets: ARID20, ARID10, and YCB10. The ARID20 and ARID10 subsets contain scenes with up to 20 and 10 objects, respectively, from the Autonomous Robot Indoor Dataset [302], while the YCB10 subset contains scenes with up to 10 objects from the Yale-CMU-Berkeley object and model set [303]. Each scene is created by adding one object at a time and capturing new frames with two ASUS-PRO cameras positioned at different heights. Scene variation is further introduced by changing the support plane and background texture. Since OCID was originally designed to evaluate object segmentation methods in cluttered scenes, semantic labels are not provided by the authors. Therefore, we adapt the dataset for a classification task by cropping out the objects from each frame, annotating them with semantic labels similar to the RGB-D Object Dataset, as we can see in Figure 2.3, and filtering out classes with fewer than 20 images. We use the crops from the ARID20 subset for training and the crops from the ARID10 subset for testing, resulting in a total of 3,939 RGB-D images capturing 49 unique objects. The original datasets, as well as the crops and annotations used in this paper, are available at https://www.acin.tuwien.ac.at/en/vision-for-robotics/software-tools/object-clutter-indoor-dataset/.

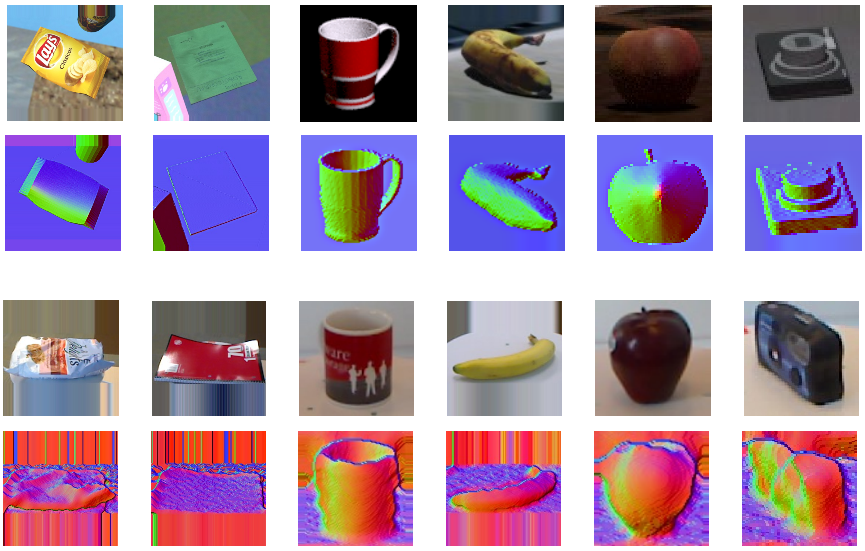


Fig. 2.4 Samples from the synthetic dataset synROD [3] and its real counterpart RGB-D Object dataset (ROD) [1]. The top two lines depict RGB and depth images from synROD, while the bottom two lines showcase the corresponding RGB and depth images from ROD, emphasizing the domain shift from synthetic to real data. The data has been processed according to the procedure outlined in [5]. The depth information is represented using surface normals, employing the best non-learned colorization method currently available.

synROD. The synROD dataset is a synthetic dataset that was created using object models from the same categories as the ROD dataset, some samples are shown in Figure 2.4. To enable comparison between the two datasets, we randomly select and extract approximately 40,000 object crops from synROD to match the dimensions of ROD. In our experiments, we evaluate domain adaptation methods by considering synROD as the synthetic source dataset and ROD as the real target dataset.

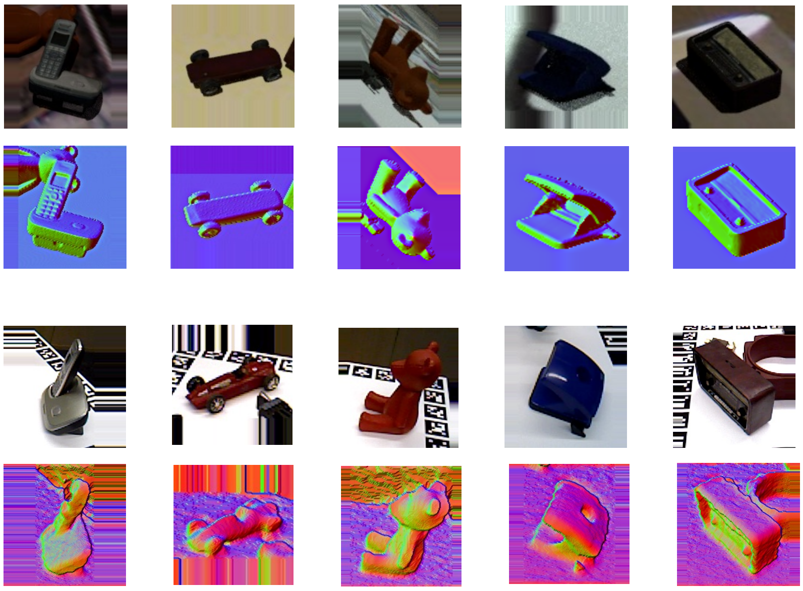
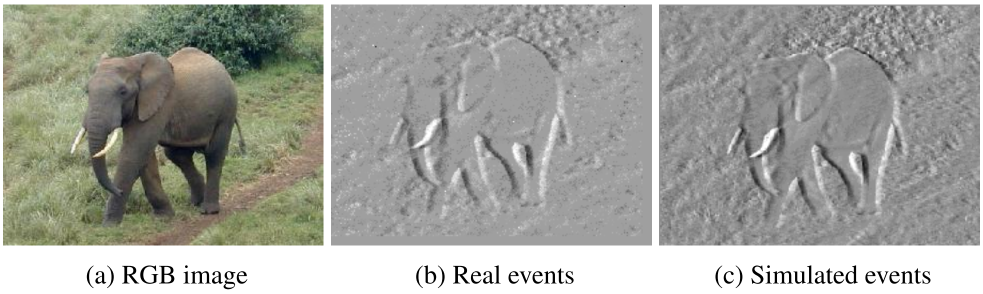


Fig. 2.5 Samples from the Homebrewed dataset [7], synthetic (synHB) and real counterpart (realHB). The top two lines display RGB and depth images from synHB, while the bottom two lines show the corresponding RGB and depth images from realHB. The data has been processed following the procedure described in [5]. Depth information is encoded using surface normals, utilizing the state-of-the-art non-learned colorization method for enhanced visualization.

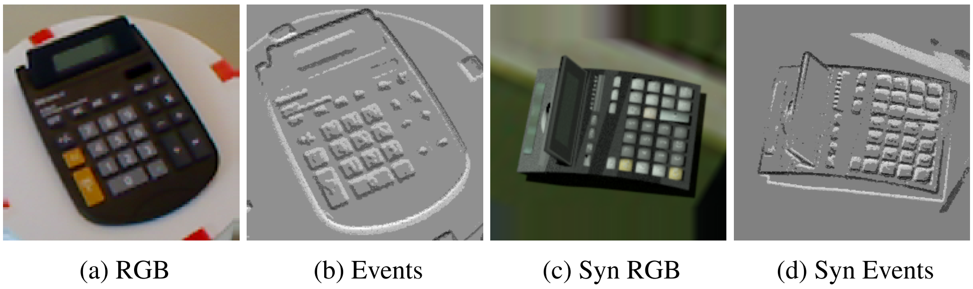
HomebrewedDB. The Homebrewed (HB) dataset [7]includes 17 toy, 8 household, and 8 industry-relevant objects, for a total of 33 instances. It provides high-quality object models reconstructed using a 3D scanner and 13 validation sequences, each of which consists of three to eight objects placed on a large turntable and recorded with two RGB-D cameras at 30° and 45° angles above the horizon. To adapt the HB dataset for the instance recognition task, we extract object crops from all the validation sequences, resulting in 22,935 RGB-D samples, which we refer to as realHB. We also create a synthetic version of the HB dataset, synHB, by rendering the reconstructed object models using the same procedure as for synROD (described above), some samples are depicted in Figure 2.5. To ensure comparability between the two datasets, we randomly select and extract approximately 25,000 object crops from synHB to match the dimensions of realHB. In our experiments, we evaluate unsupervised domain adaptation (UDA) methods by considering synHB as the synthetic source dataset and realHB as the real target dataset.



(a) RGB image (b) Real events (c) Simulated events

Fig. 2.6 Real and simulated events (voxel grid [8]) on a Caltech101 sample.

N-Caltech101. The Neuromorphic Caltech101 (N-Caltech101) dataset [23] is an event-based conversion of the Caltech-101 dataset [304].Figure 2.6 showcases a few examples from this dataset. Samples from N-Caltech101 were obtained by recording the original RGB images using an event-based camera placed in front of a still monitor displaying the images. An extension of N-Caltech101, known as the simulated N-Caltech101, was recently proposed in [294]. It was created using the ESIM simulator [44] and replicates the setup used to record the real samples. We use the simulated N-Caltech101 as the source data and the real N-Caltech101 as the target data, and follow the train and test splits provided in the EST codebase [269], and we report in our experiments the top-1 accuracy as in [294].



(a) RGB (b) Events (c) Syn RGB (d) Syn Events

Fig. 2.7 Samples from the ROD [1] dataset (a)-(b), and from the synthetic version synROD [3] (c)-(d). Event sequences are displayed using a voxel-grid [8] representation.

N-ROD. The N-ROD dataset is proposed in this work [305] and it is an extension of the popular RGB-D Object Dataset (ROD) [1]. It extends the real part (ROD) as well as the synthetic part presented above (synROD). The N-ROD dataset builds upon these components by incorporating real and simulated event recordings extracted from ROD samples, along with simulated events derived from synROD's synthetic images, few samples are reported in Figure 2.7. In this thesis, we specifically focus on the simulated part of ROD. By isolating the experiments to analyze only the simulated data, we concentrate on the problem of transferring knowledge from synthetic data to real data.

2.4.2 Action Recognition Datasets

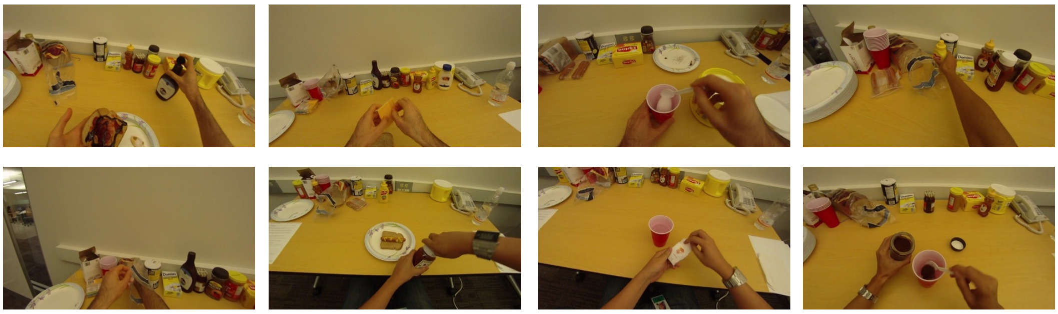


Fig. 2.8 Samples from the Georgia Tech Egocentric Activities (GTEA) dataset [9]. The videos are captured using a GoPro camera mounted on a baseball cap, which is worn by the subject. The camera is positioned to record the visual field in front of the subject's eyes, providing a first-person perspective.

In this subsection, we present details about the second group of datasets, which are specifically designed for action recognition tasks with a focus on egocentric vision. These datasets aim to capture human actions, and while not all of them are exclusively first-person, the majority incorporate egocentric videos. The modalities included in these datasets typically consist of RGB images, and additional modalities such as optical flow or audio may also be available. The availability of these datasets allows us to effectively evaluate and validate our multi-modal solutions for accurate action recognition, particularly in the context of egocentric vision.

GTEA-61. The Georgia Tech Egocentric Activities (GTEA) dataset [9] comprises 28 videos of four individuals performing seven daily activities, such as preparing sandwiches, tea, or coffee, some examples are depicted in Figure 2.8. Each video spans a one-minute recording and features approximately 20 actions. The videos were captured using a GoPro camera fixed to a baseball cap, positioned to record the area in front of the subject's eyes, resulting in a sequence of high-definition frames with a resolution of 1280x720, recorded at 30 FPS and extracted at a rate of 15 FPS, totaling 31,222 frames. To ensure comparability, we report results obtained using a fixed split (with subject S2 as the evaluation subject) and a leave-one-subject-out cross-validation setting.

Fig. 2.9 Samples from the extended GTEA Gaze+ dataset [10]. The dataset comprises 28 hours of cooking activities including 106 fine-grained action classes with an average duration of 3.2 seconds.

EGTEA Gaze+. The Extended GTEA Gaze+ dataset [10] is a comprehensive collection of first-person view (FPV) actions and gaze tracking data, which has been collected through high-definition videos accompanied by audio recordings. The dataset comprises 28 hours of cooking activities from 86 unique sessions performed by 32 subjects and includes 106 fine-grained action classes with an average duration of 3.2 seconds. The dataset provides frame-level action annotations, pixel-level hand masks at sampled frames, and gaze tracking data captured at 30Hz. With a total of 10325 instances of fine-grained actions, such as "Cutting bell pepper" and "Pouring condiment into salad," and 15,176 hand masks from 13,847 frames. Figure 2.9 showcases a few examples from this dataset. In our experiments, we present results for each of the three provided train-test splits as well as the average performance across all splits.

Fig. 2.10 Samples from the First-Person Hand Action (FPHA) benchmark dataset [11]. Frames capturing hand actions recorded in three distinct scenarios are shown.

FPHA. The First-Person Hand Action (FPHA) benchmark dataset [11] contains 1,175 action videos divided into 45 distinct action categories, which have been grouped into three scenarios: kitchen (25), office (12), and social (8). The examples shown in the Figure 2.10 provide a visual representation of the actions performed in these different scenarios. The videos were captured by 6 actors and comprise a total of 105,459 RGB-D frames that are annotated with precise hand pose and action category information. The action sequences in the dataset exhibit substantial inter-subject and intra-subject variability with regard to style, speed, scale, and viewpoint. The action categories in the dataset demonstrate a broad range of hand configurations, displaying diversity in both hand pose and action space, in line with the taxonomy presented in [306]. Each object in the dataset is associated with at least one action (e.g., 'pen-write') and a maximum of four (e.g., 'sponge-wash', 'scratch', 'squeeze', and 'flip'). These 45 hand actions were recorded in diverse settings and are representative of a broad spectrum of activities performed in daily life. To evaluate the dataset, we adopt the train/test splits proposed in [11], ensuring the consistency and comparability of the results to those obtained by other researchers.

Fig. 2.11 Samples from the EPIC-Kitchens-55 (EK55) dataset [2]. This dataset is a comprehensive and diverse collection of egocentric videos of fine-grained actions captured in the kitchens of 32 participants from 10 different countries. The videos are recorded using head-mounted cameras and are of Full HD quality.

EPIC-Kitchens-55. The EPIC-Kitchens-55 (EK55) dataset [2] is a comprehensive and varied collection of egocentric video recordings collected in the kitchens of 32 participants from 10 different countries. Figure 2.11 provides a glimpse of a few frames showcasing various actions within the dataset. The footage was captured using a head-mounted camera and is of Full HD quality, with a frame rate of 60 frames per second, yielding a total of 55 hours of recording time and 11.5 million frames. The dataset includes 39,594 annotated action segments and 454,255 object-bounding boxes around objects of interaction, along with annotations of 125 verb classes and 331 noun classes. The EK55 dataset has been evaluated using two different protocols. The first protocol follows the experimental protocol proposed in [36] by defining a custom training set (participants 1-29) and validation set (participants 30-31) to mirror the "unseen" kitchen split of the EK challenge. The second protocol, proposed in [37], involves selecting the three largest kitchens (based on the number of training action instances) to form separate domains referred to as D1, D2, and D3, corresponding to P01, P22, and P08, respectively. In this analysis, we focus on the performance of the 8 largest action classes, namely ‘put’, ‘take’, ‘open’, ‘close’, ‘wash’, ‘cut’, ‘mix’, and ‘pour’, which constitute 80% of the training action segments in these domains. This approach ensures a sufficient number of examples per domain and class without balancing the training set.

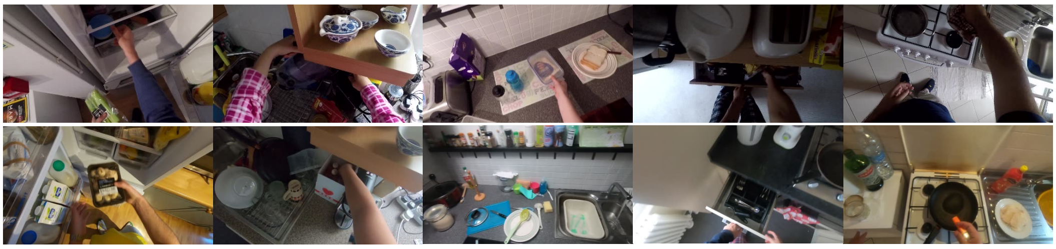


Fig. 2.12 The first row showcases samples from the EPIC-Kitchen-55 (EK55) dataset [12], while the second row features samples from the extended version, EPIC-Kitchen-100 (EK100) dataset [13]. EK100 expands upon EK55 with 100 hours of video, 20 million frames, and 90,000 actions across 700 videos. In our study, we focus on the challenging task of unsupervised domain adaptation, which involves adapting from a labeled source domain (videos recorded in 2018) to an unlabeled target domain (newly collected videos). The domain shift arises from changes in location, hardware, and temporal offsets.

EPIC-Kitchens-100. The EPIC-Kitchen-100 (EK100) dataset [13] is an extended version of the EK55 dataset, which consists of 100 hours of video, 20 million frames, and 90,000 actions across 700 videos. The videos were recorded using head-mounted cameras and depict unscripted activities in 45 environments. The new version of the dataset, EK100, is characterized by dense and comprehensive annotations of fine-grained actions, which were obtained through a new annotation pipeline. The dataset presents six challenges in the field of action recognition and action detection, including full and weak supervision, action anticipation, cross-modal retrieval from captions, and unsupervised domain adaptation. In our work, we focus on the last challenge, unsupervised domain adaptation, which involves utilizing a labeled source domain and adapting to an unlabeled target domain. The source domain consists of videos recorded in 2018, while the target domain consists of newly collected videos recorded two years later. The main causes of the domain shift include changes in location, hardware, and temporal offsets (as shown in Figure 2.12, which make this task particularly challenging.

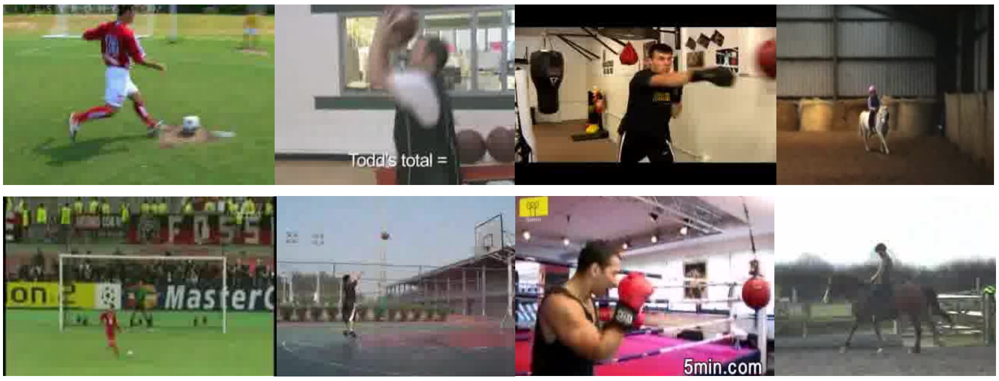


Fig. 2.13 Samples from the HMDB51 and UCF101 datasets [14, 15] . The first row showcases examples of different actions from the HMDB51 dataset, while the second row displays the corresponding action samples from the UCF101 dataset.

UCF and HMDB. The UCF101 [14] and HMDB51 [15] datasets are widely used benchmarks for human action recognition tasks. UCF101 consists of 101 action classes, containing over 13,000 video clips and 27 hours of video data. The dataset consists of user-uploaded videos with realistic conditions, including camera motion and cluttered backgrounds. It is an extension of the UCF50 dataset [307], which included 50 action classes. On the other hand, HMDB51 comprises 51 distinct action categories, with each category containing at least 101 video clips, resulting in a total of 6,766 clips. The dataset covers a wide range of sources and provides additional information such as camera viewpoint, camera motion, video quality, and the number of actors involved in the action. In the context of cross-domain analysis, a protocol described in [189] is followed. This protocol focuses on the relevant and overlapping categories between UCF101 and HMDB51, resulting in a subset of 12 common categories that are used for analysis and evaluation process.

Chapter 3

Multi-Modal Learning for Robotics Vision: Object Recognition

This chapter explores the field of object recognition in robotic vision. We thoroughly examine three key aspects of this task. Firstly, we introduce a novel method for multi-modal fusion that effectively leverages the complementary nature of different modalities, resulting in robust object recognition. In the second part, we address the challenge of data scarcity. We propose a unique technique for multi-modal adaptation that mitigates the limited availability of labeled data by allowing the utilization of synthetic data to train a model. This approach enhances the model's generalization capabilities and reduces its reliance on real annotated data. Lastly, we explore the cutting-edge sensor known as an event camera. We investigate its potential in object recognition tasks and evaluate its performance compared to traditional visual sensors. By exploring these areas, this chapter aims to contribute to the advancement of robotic vision by introducing multi-modal object recognition techniques.

3.1 Multi-Modal Fusion – RCFusion

Despite the significant progress made in the field of robotics vision, there is still a challenge in effectively leveraging both RGB and depth data in a synergistic manner to enhance object recognition. In this section, we describe a novel end-to-end architecture for RGB-D object recognition, which we refer to as Recurrent Convolutional Fusion (RCFusion). Our architecture combines RGB and depth information at various abstraction levels to create concise, highly discriminatory multi-modal features. This distinguishes our approach from prior research that primarily focuses on adapting depth images to reuse pre-trained RGB CNNs, utilizing colorization techniques. Experimental results were obtained using two widely used datasets, the RGB-D Object Dataset and JHUIT-50. These results demonstrate that RCFusion outperforms state-of-the-art techniques in both object categorization and instance recognition tasks. Further experiments on the more challenging Object Clutter Indoor Dataset confirm the validity of our method in cluttered and occluded environments.

Human environments are comprised of objects, and the successful completion of daily activities necessitates a comprehensive understanding and manipulation of these objects. Robotic systems designed to assist in such environments must exhibit exceptional object recognition capabilities. This recognition represents the foundation for higher-level tasks that critically depend on an accurate visual representation of the environment. While object recognition techniques utilizing standard RGB images have produced noteworthy results, these methods suffer from the limitations arising from projecting the three-dimensional world into a two-dimensional image plane. The integration of range imaging technologies in RGB-D (Kinect-style) cameras has the potential to address these limitations by providing a depth image that conveys information regarding the distance between the camera and the scene. The widespread adoption of RGB-D cameras in robotics is attributed to their affordability and the richness of visual information they provide. The RGB image holds information related to color, texture, and appearance, while the depth image comprises additional geometric information and demonstrates greater resilience to lighting and color variations. Since RGB-D cameras are already deployed in most service robots, improving the performance of robot perceptual systems through better integration of RGB and depth information would constitute a “free lunch". Following the groundbreaking work of Krizhevsky *et al.* [308], deep convolutional neural

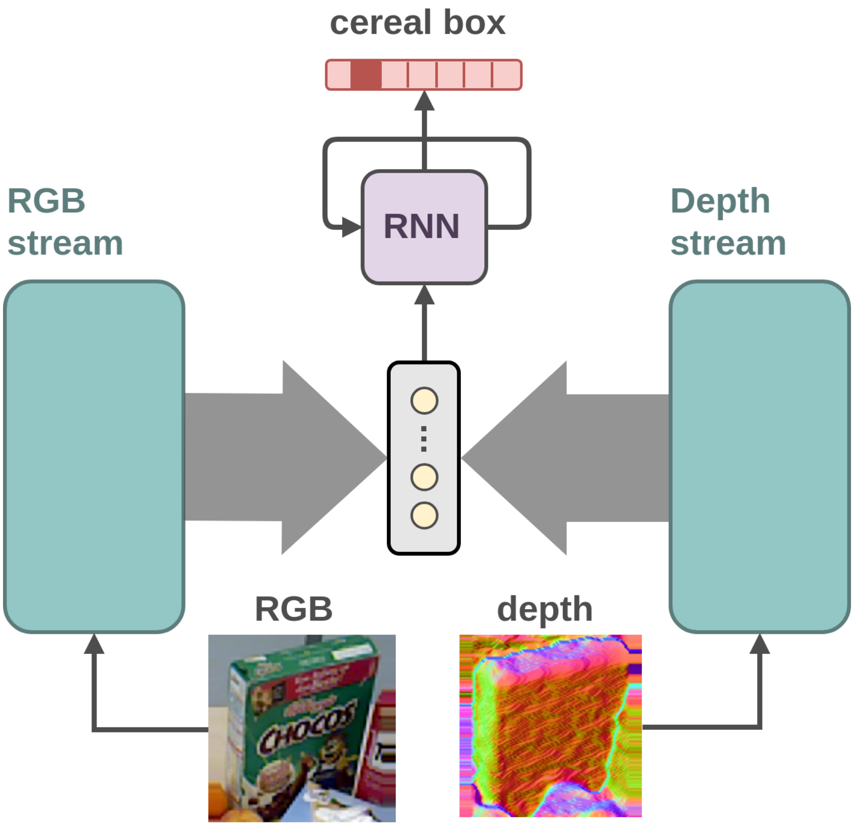


Fig. 3.1 High-level scheme of RCFusion. The blue boxes are instantiated with convolutional neural networks and the thick arrows represent multiple feature vectors extracted from different layers of a CNN.

networks (CNNs) rapidly became the dominant approach in the field of computer vision, surpassing previous state-of-the-art results across a broad range of tasks. The trend of utilizing CNNs in RGB-D object recognition was similarly observed, with several algorithms (e.g. [309, 5, 310]) relying on features learned from CNNs rather than the conventional hand-crafted features. The typical approach consists of two separate CNN streams, one operating on the RGB image and the other on the depth image, serving as feature extractors. However, the absence of a large-scale depth image dataset for training the depth CNN necessitated the development of alternative methods. A significant amount of effort has been invested in coloring depth images to enable the use of CNNs pre-trained on RGB images [42]. Nonetheless, the strategies for combining the features extracted from both modalities have been neglected. Many methods simply extract features from a specific layer of both CNNs and combine them through a fully connected or max pooling layer. Our contribution sheds light on the suboptimality of existing strategies in fusing RGB and depth information due to two underlying assumptions. Firstly, the assumption is that the selected layer always represents the optimal abstraction level for the fusion process. Secondly, the assumption that the full range of information from both modalities is not exploited during the fusion process.

In this section, we introduce the novel end-to-end architecture for RGB-D object recognition called recurrent convolutional fusion (RCFusion). Our method involves the extraction of features from multiple hidden layers of CNNs for both the RGB and depth modalities. Subsequently, the extracted features are fused via a Recurrent Neural Network (RNN), as depicted in Figure 3.1. Our hypothesis is that the fusion of features from multiple levels of abstraction from both modalities will provide more robust and discriminative information for object recognition. Although RNNs are commonly used for processing sequential data, they have also been demonstrated to be highly effective information compression mechanisms [311]. Additionally, RNNs are able to effectively scale in terms of parameters with regard to the number of extracted features.

The results of our experiments demonstrate that the proposed RCFusion architecture outperforms the baseline approach, which uses a fully connected layer to combine the features from both modalities. The proposed method establishes new state-of-the-art results on two standard object recognition benchmarks: the RGB-D Object Dataset [16] and JHUIT-50 [6]. In addition, to further consolidate the effectiveness of our method, we modify the Object Clutter Indoor Dataset (OCID)[17], an object segmentation dataset, for the instance recognition task. OCID was recently introduced to provide scenes with high levels of clutter and occlusion, which pose significant challenges for robotic visual perception systems[302]. Despite the limited amount of training data, our method proves to be effective even on this challenging dataset.

In summary, our contributions are:

- A new architecture for RGB-D object recognition that fuses RGB and depth features from multiple levels of abstraction in a sequential manner.
- State-of-the-art performance on two popular RGB-D object recognition benchmark datasets.
- Introduction of a new benchmark with robotic-oriented challenges.

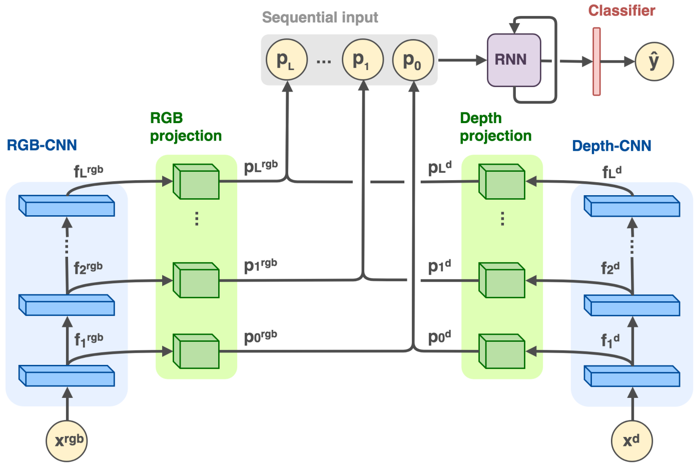


Fig. 3.2 The architecture of Recurrent Convolutional Fusion (RCFusion). It consists of two parallel streams of Convolutional Neural Networks (CNNs) that process RGB and depth images respectively. The features extracted from the corresponding hidden layers of the two CNNs are then projected into a shared representation space, concatenated, and sequentially fed into a Recurrent Neural Network (RNN) for synthesis. The final multi-modal features produced by the RNN are then utilized by a classifier to predict the label of the input data.

The rest of the section is organized as follows. Section 3.1.1 describes the proposed RCFusion method in detail. Section 3.1.2 provides the experimental results to support the effectiveness of RCFusion, and Section 3.1.3 summarizes the findings and highlights the implications of the study.

3.1.1 Recurrent Convolutional Fusion

The proposed RGB-D Object Recognition approach, referred to as RCFusion, is depicted in Figure 3.2. This multi-modal architecture consists of three key stages:

1. *multi-Level feature extraction:* the system comprises two convolutional neural networks - RGB-CNN and Depth-CNN, which are used to process RGB and depth data, respectively. Both of these networks share the same underlying architecture and are employed to extract features from different levels of the network.

2. *feature projection and concatenation:* the features extracted from each level of RGB-CNN and Depth-CNN are transformed using projection blocks and concatenated to generate the corresponding RGB-D feature.

3. *recurrent multi-modal fusion:* the RGB-D features obtained from the previous stage are fed sequentially to an RNN that generates a compact and descriptive multi-modal feature.

Finally, the output of the RNN is utilized by a softmax classifier to predict the object label. The network can be trained in an end-to-end manner using standard backpropagation algorithms and cross-entropy loss based on stochastic gradient descent. The following sections will provide a more in-depth explanation of the aforementioned aspects of RCFusion.

Multi-level Feature Extraction CNNs are often used in computer vision to process input data using sets of filters learned from large amounts of data. These filters represent progressively higher levels of abstraction, from edges and textures to patterns, parts, and objects [312]. In RGB-D object recognition, it is common practice to combine the output of one of the last layers of the RGB-CNN and Depth-CNN, typically the last layer before the classifier, and assume that this layer represents the appropriate level of abstraction to combine the two modalities. However, we argue that it is possible to remove this assumption by combining RGB and depth information at multiple layers across the CNNs, and use them all to generate a highly discriminative RGB-D feature.

Let $x^{rgb} \in \mathcal{X}^{rgb}$ denote the RGB input images, $x^{d} \in \mathcal{X}^{d}$ the depth input images, and $y \in \mathcal{Y}$ the labels, where $\mathcal{X}^{rgb}$, $\mathcal{X}^{d}$, and $\mathcal{Y}$ are the RGB/depth input and label spaces. We also denote the output of layer i of RGB-CNN and Depth-CNN as f_i^{rgb} and f_i^{d}, respectively, with $i = 1,...,L$ and L being the total number of layers in each CNN. Notably, visualization of learned filters has shown that for a given task, a chosen CNN architecture consistently generates features with the same level of abstraction from a reference layer [312]. For example, the AlexNet architecture [308] learns various types of Gabor filters in the first convolutional layer. To ensure the same level of abstraction at corresponding layers, we, therefore, employ the same architecture for both RGB- and Depth-CNNs.

Feature Projection and Concatenation Combining features from different hidden layers of a network presents a significant challenge due to the lack of a one-to-one correspondence between elements of the different feature vectors. Specifically, the feature vectors f_i^{*} and f_j^{*}, where $i \neq j$ and $*$ represents either *rgb* or *d*, generally have

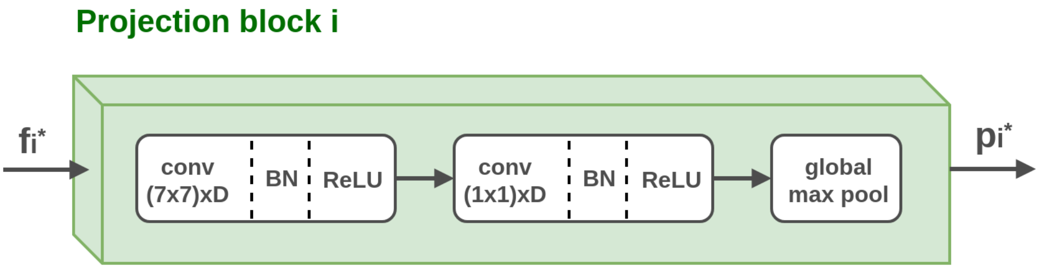


Fig. 3.3 Projection block. It converts the input feature f_i^* into a projected feature p_i^*. The block consists of a convolutional layer $conv(k \times k) \times D$ with D filters of size $(k \times k)$, followed by a batch normalization layer (BN) and an activation layer with ReLU non-linearity ($ReLU$).

different dimensions and belong to distinct feature spaces, $\mathfrak{F}_i$ and $\mathfrak{F}_j$, respectively. In order to enable comparison between features obtained from different levels of abstraction, we project them into a common space $\bar{\mathfrak{F}}$ using the projection block $G_i(.)$ defined by

$$p_i^* = G_i^*(f_i^*) \quad \text{s.t.} \quad p_i^* \in \bar{\mathfrak{F}} \tag{3.1}$$

Here, $G_i(.)$ is a non-linear transformation that maps a volumetric input into a vector of dimensions (1×D) using two convolutional layers with batch normalization and ReLU non-linearity, as well as a global max pooling layer, as illustrated in Figure 3.3. The projected RGB and depth features of each layer i are concatenated to form $p_i = [p_i^{rgb}; p_i^d]$.

Recurrent Multi-Modal Fusion In order to create a compact multi-modal representation, the set of projected features$\{p_1, \ldots, p_L\}$ is sequentially fed to an RNN, which aligns the positions of the sequence elements to steps in computation time and generates a sequence of hidden states h_i as a function of the previous hidden state h_{i-1} and the current input p_i. Specifically, we adopt the gated recurrent unit (GRU)[313] as the RNN, which is a variation of the long-short term memory (LSTM)[314] that requires 25% fewer parameters. GRU has been shown to retain information even in extremely long sequences with thousands of elements [311].

GRU computes the n^{th} element of the hidden state at step i as an adaptive linear interpolation:

$$h_i^n = (1 - z_i^n)h_{i-1}^n + z_i^n \tilde{h}_i^n, \tag{3.2}$$

where z_i^n is called update gate and is computed as

$$z_i^n = sigmoid(\theta_z p_i + \gamma_z h_i)^n, \tag{3.3}$$

where $sigmoid(.)$ is the sigmoid function and θ_z and γ_z are the trainable parameters of the gate. Essentially, the update gate determines how much the unit updates its content. The candidate activation $\tilde{h}_i$ in Equation 3.2 is computed as

$$\tilde{h}_i^n = tanh(\theta_h p_i + \gamma_h (r_i \odot h_{i-1}))^n, \tag{3.4}$$

where r_i is the reset gate, θ_h and γ_h are trainable parameters and $\odot$ is the element-wise multiplication operation. Similarly to z_i^n, the reset gate r_i^n is computed as

$$r_i^n = sigmoid(\theta_r p_i + \gamma_r h_i)^n, \tag{3.5}$$

where θ_r and γ_r are the trainable parameters of the gate. The purpose of the reset gate is to reset the hidden state of the network to the current input p_i when r_i^n assumes values close to zero. This double-gate mechanism is designed to ensure that the hidden state captures the most relevant information of the input sequence $\{p_1, \ldots, p_L\}$. in a progressive manner.

The RNN, together with a softmax classifier, serves as a parametric function that models a probability distribution over a given sequence, by learning to predict the class label given the sequence of projected RGB-D features. Specifically, the prediction of the j^{th} class of the multinomial distribution of K object categories can be obtained as follows

$$\hat{y}_j = Pr(y_j = 1 | p_1, ..., p_1) = \frac{exp(h_L^T \theta_c^j)}{\sum_{k=1}^K exp(h_L^T \theta_c^k)}, \tag{3.6}$$

where θ indicates the matrix of trainable parameters of the classifier and $\theta_{j(/_k)}$ represents its $j^{th}(/k^{th})$ row, and T represents the transpose operation.

The use of a recurrent neural network for this task serves two purposes in our approach. Firstly, the hidden state of the network acts as a memory unit and effectively summarizes the most relevant information from the different levels of abstraction. Secondly, the number of trainable parameters in the network is independent of the length of the input sequence, unlike a more traditional choice, such as a fully connected layer, where the number of parameters grows linearly with the sequence length. While RNNs are commonly used for processing time series data, previous works [315, 316] have demonstrated their effectiveness in compressing and combining information from various sources.

3.1.2 Experiments

In this section, we present an evaluation of RCFusion using three datasets: RGB-D Object Dataset, JHUIT-50, and OCID, which were previously introduced in section 2.4.1. We begin by disclosing the experimental protocol that was employed and subsequently describing the network training settings. We then proceed to compare the performance of our method with that of existing approaches. Finally, an ablation study is conducted to ascertain the contributions of individual components of our method.

Architecture The architecture of RCFusion comprises three principal components: RGB-/Depth-CNN, projection blocks, and RNN, each with unique design choices.

RGB-/Depth-CNN: To ensure computational and memory efficiency, we employ a CNN architecture with a relatively small number of parameters. We adopt ResNet-18, which is the most compact version proposed by He *et al.* [317], in accordance with the commonly used residual network architecture. ResNet-18 comprises 18 convolutional layers organized into five residual blocks, with approximately 40,000 parameters. We extract features after each of the network's two skip connections per residual block, beginning from the second block, resulting in $L = 8$ extracted features per network. An implementation of pre-trained ResNet-18 on ImageNet is available in [318].

Projection blocks: The projection blocks, illustrated in Figure 3.3, are designed to exploit firstly the spatial dimensions of the input, width and height, using the first convolutional layer with $D = 512$ filters of size (7×7). Then, with the second convolutional layer with $D = 512$ filters of size (1×1) focuses on the depth. Finally, using global max pooling, the maximum of each depth slice is computed. Among the various configurations we attempted, this particular instantiation of the projection blocks yielded the best performance.

RNN: To balance the capacity of the network with the limited number of parameters, we adopt the commonly used GRU [313]. In our experiments, we employ a single GRU layer with $M = 50$ memory neurons to process the sequence of projected vectors. GRU can be implemented using various deep learning libraries, including TensorFlow.

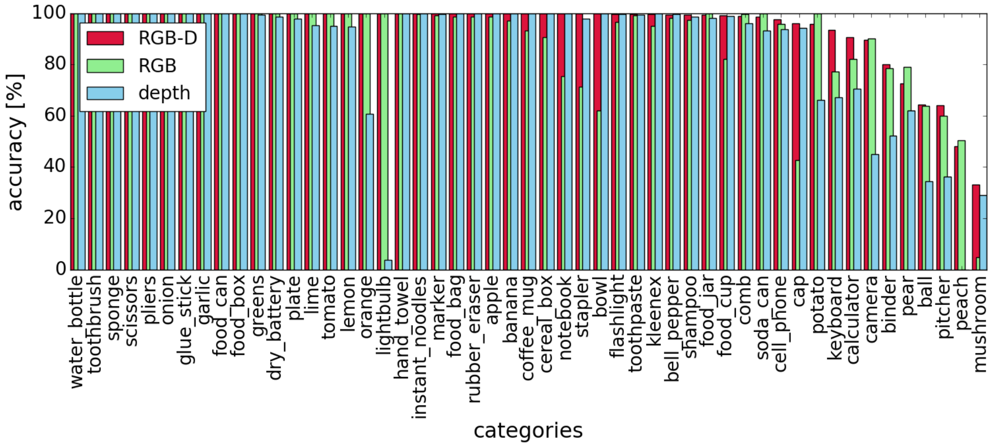


Fig. 3.4 Per class accuracy (%) of RCFusion on RGB-D Object Dataset [16].

Training RCFusion model is trained using the RMSprop optimizer with a batch size of 64, a learning rate of 0.001, momentum of 0.9, weight decay of 0.0002, and a maximum norm of 4. We set the projection depth parameter to $D = 512$ and the number of memory neurons to $M = 50$ through a grid search. The weights of the two ResNet-18 networks used for the RGB- and Depth-CNN are initialized with values obtained through pre-training on ImageNet, while the rest of the network is initialized using the Xavier initialization method in a multi-start fashion. During training, all parameters of the network, including those that define the RGB- and Depth-CNN, are updated. The input to the network is synchronized RGB and depth images, preprocessed according to the procedure described in [5]. Depth information is encoded using surface normals, which is currently the best non-learned colorization method available. To compensate for the small training set sizes of JHUIT-50 and OCID, we employ simple data augmentation techniques, including scaling, horizontal and vertical flipping, and 90 degree rotation.

Results To demonstrate the effectiveness of our method, we conduct a thorough evaluation on multiple benchmark datasets. Specifically, we first compare the performance of RCFusion with state-of-the-art methods on two commonly used datasets, RGB-D Object Dataset and JHUIT-50 We then challenge the robustness of our method on a more difficult dataset, OCID, and conduct an ablation study to highlight the contribution of each component of our approach.

RGB-D OBJECT DATASET			
Method	RGB	Depth	RGB-D
LMMMDL [319]	74.6±2.9	75.5.8±2.7	86.9±2.6
FusionNet [309]	84.1±2.7	83.8±2.7	91.3±1.4
CNN w/ FV [320]	**90.8**±1.6	81.8±2.4	93.8±0.9
DepthNet [310]	88.4±1.8	83.8±2.0	92.2±1.3
CIMDL [321]	87.3±1.6	84.2±1.7	92.4±1.8
FusionNet enhenced [5]	89.5±1.9	84.5±2.9	93.5±1.1
DECO [33]	89.5±1.6	84.0±2.3	93.6±0.9
RCFusion	89.6±2.2	**85.9**±2.7	**94.4**±1.4

Table 3.1 Top-1 accuracy (%) achieved by various object recognition methods on RGB-D Object Dataset [16]. The highest result is highlighted in **bold**, while other notable results are underlined.

JHUIT-50			
Method	RGB	Depth	RGB-D
DepthNet [310]	88.0	55.0	90.3
FusionNet enhenced [5]	94.7	56.0	95.3
DECO [33]	94.7	**61.8**	95.7
RCFusion	**95.1**	59.8	**97.7**

Table 3.2 Top-1 accuracy (%) achieved by various object recognition methods on the JHUIT-50 dataset [6]. The highest result is highlighted in **bold**, while other notable results are underlined.

Benchmark: In Table 3.1, we present the results on RGB-D Object Dataset for the object categorization task, where we train a classifier on the final features of the RGB- and Depth-CNN for each modality. Our method achieves the best multi-modal RGB-D results, outperforming all the competing approaches. Notably, the single modality predictions demonstrate that ResNet-18 is a valid trade-off between a small number of parameters and high accuracy. On the RGB modality, ResNet-18 achieves the second-highest accuracy, only behind [320], who use a VGG network [322] with considerably more parameters. For the depth modality, ResNet-18 provides higher accuracy than all the competing methods, demonstrating the effectiveness of our approach.

MISCLASSIFICATION CASES			
Reference class	RGB	Depth	RGB-D
calculator	keyboard	hand towel	hand towel
keyboard	calculator	binder	calculator
pear	apple	apple	apple
potato	lime	lime	lime

Table 3.3 Analysis of the most frequently misclassified classes in RGB, depth, and RGB-D modalities with respect to selected reference classes.

To gain a deeper understanding of the performance of RCFusion, we examine the accuracy of the model on individual categories in RGB-D Object Dataset. Figure 3.4 reveals that our multi-modal approach either matches or outperforms the results of single modalities for nearly all categories. For categories where the accuracy of one modality is extremely low, such as "lightbulb," "orange," or "bowl," RCFusion learns to rely on the other modality. An insightful analysis of the method's operation is given by comparing, for each category, which other categories cause misclassification. Table 3.3 displays, for a few sample classes, the most frequently misclassified class in the RGB, depth, and RGB-D cases. When an object class is confused with different classes in the individual modalities, as in the case of "keyboard" and "calculator," the RGB-D modality can perform better. However, when an object class is confused with the same classes in both RGB and depth modalities, such as "pear" and "potato," the RGB-D modality's performance may be slightly worse than that of single modalities. This finding reveals a weakness of the method that we plan to investigate further in the future.

Table 3.2 presents the results of RCFusion compared to other approaches on the JHUIT-50 dataset for the instance recognition task. The performance of ResNet-18 is again remarkable for the individual modalities. For the multi-modal RGB-D classification, our method significantly outperforms all the existing approaches, including the best performing method, DECO [33], by a margin of 2%. In conclusion, RCFusion sets new state-of-the-art results on the two most popular datasets for RGB-D object recognition, demonstrating its ability to perform well across different datasets and tasks.

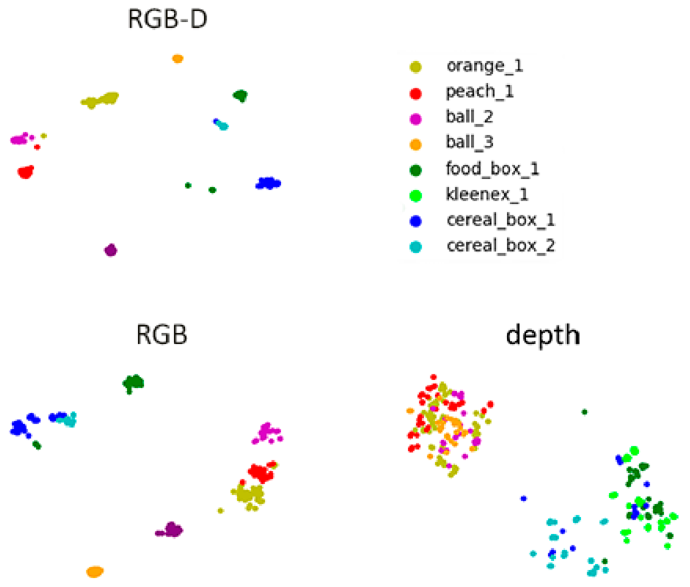


Fig. 3.5 The figure presents a t-SNE visualization of features extracted from three modalities (RGB, Depth, and RGB-D) in object recognition.

Challenge: To assess the effectiveness of our approach in challenging scenarios, we present experiments on the OCID dataset. This dataset was specifically designed to include heavily cluttered and occluded object scenes (as shown in Figure 3.6), making it particularly relevant to evaluate algorithms for RGB-D object recognition. Due to the ambiguous views presented in clutter, using multiple modalities is essential to improve recognition performance. In addition, the small training set size of 2,428 cropped images poses an additional challenge. For the instance recognition task, we report the results on OCID in Table 3.4, alongside the performance of DECO, a method that demonstrated competitive performance on RGB-D Object Dataset and JHUIT-50. The results on the single modalities reveal that depth data alone is not informative enough for this task, with an accuracy gap of 50% compared to the RGB modality. However, our proposed method effectively leverages both modalities, achieving a 6.1% improvement in accuracy over the RGB modality alone. In contrast, DECO's performance remains the same as the RGB modality, even in the multi-modal case, due to its simplistic modality fusion strategy, which selects the class with the highest probability among the RGB and depth predictions. Our more complex fusion strategy in RCFusion leads to a significant improvement of over

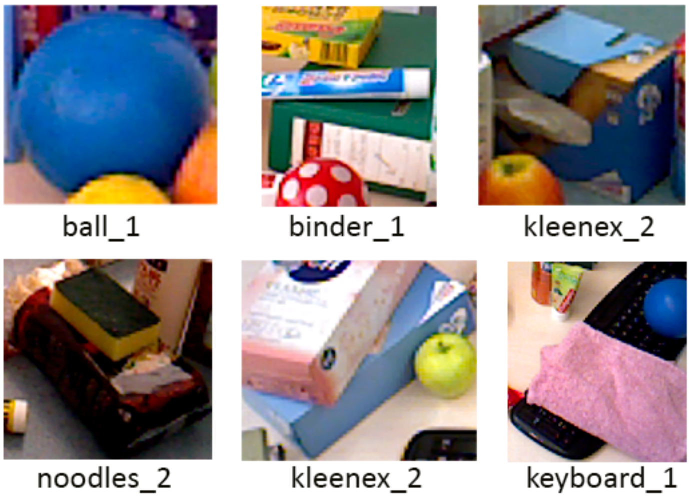


Fig. 3.6 Object crops with their respective instance labels obtained from the Object Cluttered Indoor Dataset [17]. The figure shows a selection of examples from the dataset, highlighting the diversity of objects and scenes included in the dataset.

10% in accuracy compared to DECO, highlighting the superiority of our approach in challenging scenarios.

Feature analysis: The effectiveness of RCFusion can be intuitively understood through feature analysis of the OCID dataset. Figure 3.5 presents a two-dimensional t-SNE embedding of the final features obtained from different modalities. The depth features tend to cluster together objects with similar shapes, such as those with near-spherical shapes like “orange_1", “pear_1", and “ball_2(/3)". As expected, the RGB modality offers more discriminative features, but it fails to distinguish similar pairs of objects, such as (“orange_1"-“peach_1") and (“cereal_box_1"-“cereal_box_2") that are located very close to each other. In contrast, the RGB-D features effectively separate each object into distinct clusters.

Ablation study: In order to evaluate the individual contributions of the multi-level feature extraction and recurrent fusion elements of RCFusion, we conducted an ablation study by removing each of these elements and comparing the resulting performance with the full version of the method. Table 3.4 presents the results of these variations on the OCID dataset. We observed that using only the features from the last layer of the RGB-/Depth-CNN (RCFusion - res5) caused a drop in accuracy of 2%. This confirms that utilizing features from multiple levels of abstraction leads

Object Clutter Indoor Dataset			
Method	RGB	Depth	RGB-D
DECO [33]	80.7	**36.8**	80.7
RCFusion	**85.5**	35.0	**91.6**
RCFusion - res5	-	-	89.6
RCFusion - fc	-	-	88.5

Table 3.4 Top-1 accuracy (%) obtained by DECO [33] and variations of the RCFusion method on the Object Clutter Indoor Dataset [17]. The RCFusion method is modified in two ways, with “RCFusion - res5" referring to the variation that only uses features from the last residual layer (res5) for classification and “RCFusion - fc" referring to the variation that uses a fully connected layer instead of a recurrent neural network to combine RGB and depth features. The highest result is highlighted in **bold**, while other notable results are underlined.

to better multi-modal recognition compared to relying solely on the final features of individual modalities. Similarly, when we concatenated the multi-modal features from the projection blocks and fused them with a fully connected layer instead of using the RNN, the performance dropped by 3.1% in accuracy. This highlights the importance of employing a more sophisticated fusion mechanism that effectively combines the modalities while preserving the crucial information from different levels of abstraction to obtain a discriminative RGB-D feature.

3.1.3 Discussion and Conclusion

In this section, we introduce RCFusion: a multi-modal deep neural network for RGB-D object recognition. The method comprises two streams of convolutional networks that extract RGB and depth features from multiple levels of abstraction. These features are concatenated and sequentially fed to a recurrent neural network (RNN) to obtain a compact RGB-D feature used by a softmax classifier for final classification. Our approach outperforms existing methods for RGB-D recognition on two standard benchmarks, RGB-D Object Dataset and JHUIT-50, demonstrating the validity of our approach. We also evaluate RCFusion on OCID, a challenging dataset with highly cluttered and occluded scenes and few training samples. Despite these challenges, our approach achieves compelling results and demonstrates the superiority of our multi-modal fusion.

A potential avenue for future research could be to investigate the weakness of RCFusion that was identified in our experiments. Specifically, when an object class is confused with the same classes in both RGB and depth modalities (e.g., "pear" and "potato"), the performance of the RGB-D modality may be slightly worse than that of the individual modalities. To address this limitation, we will explore the use of an ensemble of different fusion mechanisms to improve recognition accuracy for such cases.

Other interesting research in future work, we plan to explore more complex configurations of recurrent networks for multi-level sequential fusion and investigate the potential of extending the proposed approach to more complex architectures such as transformer-based models. The implementation-agnostic nature of our approach also opens up possibilities for adaptation to different tasks. The promising results obtained on object categorization further encourage the extension of this approach to higher-level tasks, such as object detection and semantic segmentation.

3.2 Multi-Modal Alignment – RR

In this section, we explore a solution to address the challenges of applying multi-modal algorithms in real-world scenarios by mitigating the issues that can arise due to the use of synthetic datasets. Specifically, this study focuses on the Unsupervised Domain Adaptation (DA) technique, which utilizes the supervision of a label-rich source dataset to make predictions on an unlabeled real-target dataset by aligning the two data distributions. Current DA methods are not well-suited to handle the multi-modal nature of RGB-D data, which is extensively used in robotic vision. To overcome this limitation, we present a novel RGB-D DA method that addresses the synthetic-to-real domain shift by utilizing the inter-modal relationship between RGB and depth images. The proposed method involves training a convolutional neural network to solve the main recognition task as well as the pretext task of predicting the relative rotation between the RGB and depth image. To assess the effectiveness of this approach and promote further research in this area, two benchmark datasets are introduced for object categorization and instance recognition. Through extensive experiments, the study demonstrates the advantages of leveraging inter-modal relations for RGB-D DA.

In section 3.1, we discussed the importance of using a multi-modal object recognition model for robotic systems to be able to understand and interact with their environment. However, a significant challenge in deploying such models in robotics is the high cost of acquiring a large amount of annotated data needed for training. To address this challenge, a promising approach that requires no manual annotation involves generating a synthetic training dataset using computer graphics software like Blender [4] to create 3D object models. However, the discrepancy between the synthetic (source) training data and the real (target) test data significantly undermines the recognition performance of the CNN.

Unsupervised Domain Adaptation (UDA) is a critical area of research in computer vision that addresses the challenges associated with transferring knowledge from a source domain to a target domain. This is accomplished by treating the source and target data as originating from different marginal distributions. UDA methods enable us to predict target samples using only annotated source samples, while unlabeled target samples are used transductively. In recent years, significant progress has been made in the development of UDA techniques aimed at reducing the domain shift between the source and target distributions. These methods operate at both the

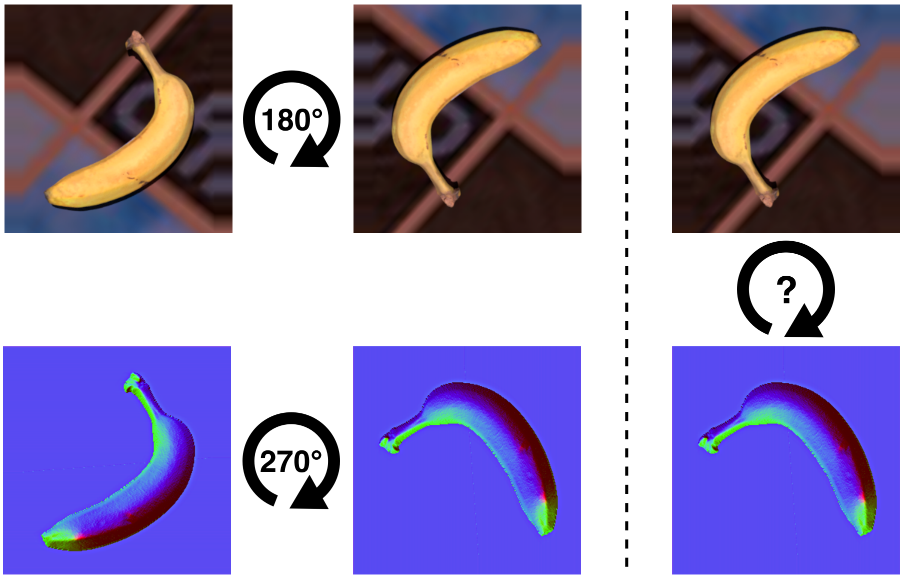


Fig. 3.7 Q: “By how much should the RGB image (top) be rotated to align with the depth image (bottom)?" A: “90°". This question describes the self-supervised task of predicting the relative rotation between the RGB and depth image (shown with surface normal colorization [18]) of a sample after they have been independently rotated.

feature level [149, 154] and the pixel level [156, 323], improving the generalization performance of the model on the target domain. Despite these advances, existing DA strategies assume that the data arise from a single modality, and consequently, may result in sub-optimal performance when confronted with multi-modal data. The natural inter-modal relationships among the data are ignored in such cases, leading to a loss of valuable information.

This section presents a novel domain adaptation (DA) method specifically designed for RGB-D data. Our proposed approach involves solving a multi-task learning problem, where a convolutional neural network (CNN) is trained to solve a supervised main task, which is object recognition, and a self-supervised pretext (or auxiliary) task from pairs of RGB and depth images. To encourage the network to generate domain-invariant features, we create an artificial problem by rotating the RGB and depth image of a sample and asking the network to predict the relative rotation that re-aligns them, as shown in Figure 3.7. Due to the self-supervised nature of this pretext task, we can use both source and target data to train the model, while the supervision of the source data is used to train the model on the main task, as depicted in Figure 3.8. To assess the performance of our method on object categorization and instance recognition, we introduce two benchmark datasets, each consisting of a

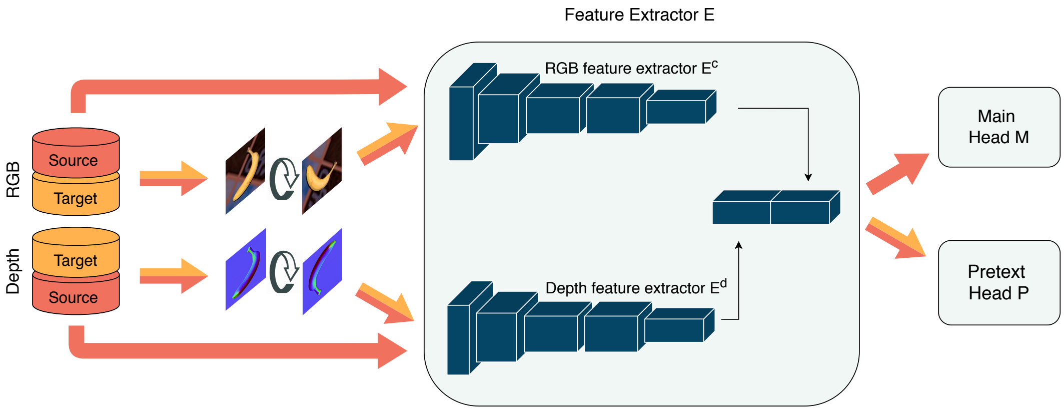


Fig. 3.8 Overview of our method for RGB-D domain adaptation, which uses a convolutional neural network (blue squares) consisting of a two-stream feature extractor E and two network heads. The main head M is trained for object recognition using the labeled source data (red arrow), while the pretext head P is trained using both source and target data (orange+red arrow), with independent rotation of RGB and depth images .

synthetic and a real part. Our proposed approach demonstrates promising results, highlighting its potential for effective DA in the context of RGB-D data. For instance recognition, we utilize the HomeBrewedDB (HB)[7] models as the source dataset and real RGB-D sequences of the same dataset as the target dataset. For object categorization, no dataset comprises both synthetic and real data. Therefore, we use the popular RGB-D Object Dataset (ROD)[324] for the real data and create the synthetic counterpart ourselves. To this end, we introduce synROD: a dataset generated by collecting and rendering 3D object models from the same categories as ROD using publicly available Web resources. Our extensive experiments on these datasets demonstrate that our proposed pretext task effectively reduces the synthetic-to-real domain gap and outperforms existing DA methods that do not leverage the inter-modal relations of RGB-D data.

In summary the contribution of this section are:

- a novel multi-modal DA algorithm for RGB-D object recognition that reduces the domain gap by leveraging the relation between RGB and depth data,
- two benchmark datasets for evaluating RGB-D DA methods on object categorization and instance recognition (including the newly collected synROD), and

- quantitative and qualitative experiments demonstrating the superior performance of our method compared to existing DA approaches.

The rest of the paper is organized as follows: the next section 3.2.1 describes synROD, section 3.2.2 introduces the proposed method, section 3.2.3 presents the experimental results and section 3.2.4 draws the conclusions.

3.2.1 Dataset

In this section, we describe synROD and the protocol followed for its creation. More specifically, the next section describes the criteria used to define the scope of the dataset and collect the 3D object models from Web resources; and then we illustrate the procedure used to render 2.5D scenes from the 3D object models. All the collected data, together with the information needed to replicate the experiments is publicly available at “https://www.acin.tuwien.ac.at/en/vision-for-robotics/software-tools/synthetic-to-real-rgbd-datasets/".

Selecting 3D Object Models RGB-D DA has not been thoroughly explored in the literature, resulting in the absence of standardized benchmark datasets to evaluate newly developed methods. A major challenge in creating a suitable dataset for evaluating DA techniques is to identify two distinct sets of data that share the same annotated classes but have been acquired under different conditions. In particular, we are interested in the synthetic-to-real domain shift, where the source domain presents RGB-D synthetic data, while the target domain presents RGB-D real data. Existing 3D object datasets, such as ModelNet [325] and ShapeNet [326], do not have a corresponding real dataset that shares the same classes. Moreover, some models lack texture, making them unsuitable for the study’s purpose, which requires both object shape (depth) and texture (color). On the other hand, other available datasets, such as LineMOD [327] and HB, offer both real object data and reconstructed 3D models. However, these datasets lack category-level annotation, making them appropriate only for instance recognition, not object categorization.

To overcome this problem, we collect a new synthetic dataset called synROD. We selected the object models for synROD in such a way that each one belongs to one of the 51 categories defined by ROD, the most used RGB-D dataset in robotics for object categorization [61, 18]. We query the objects from the free catalogs of public

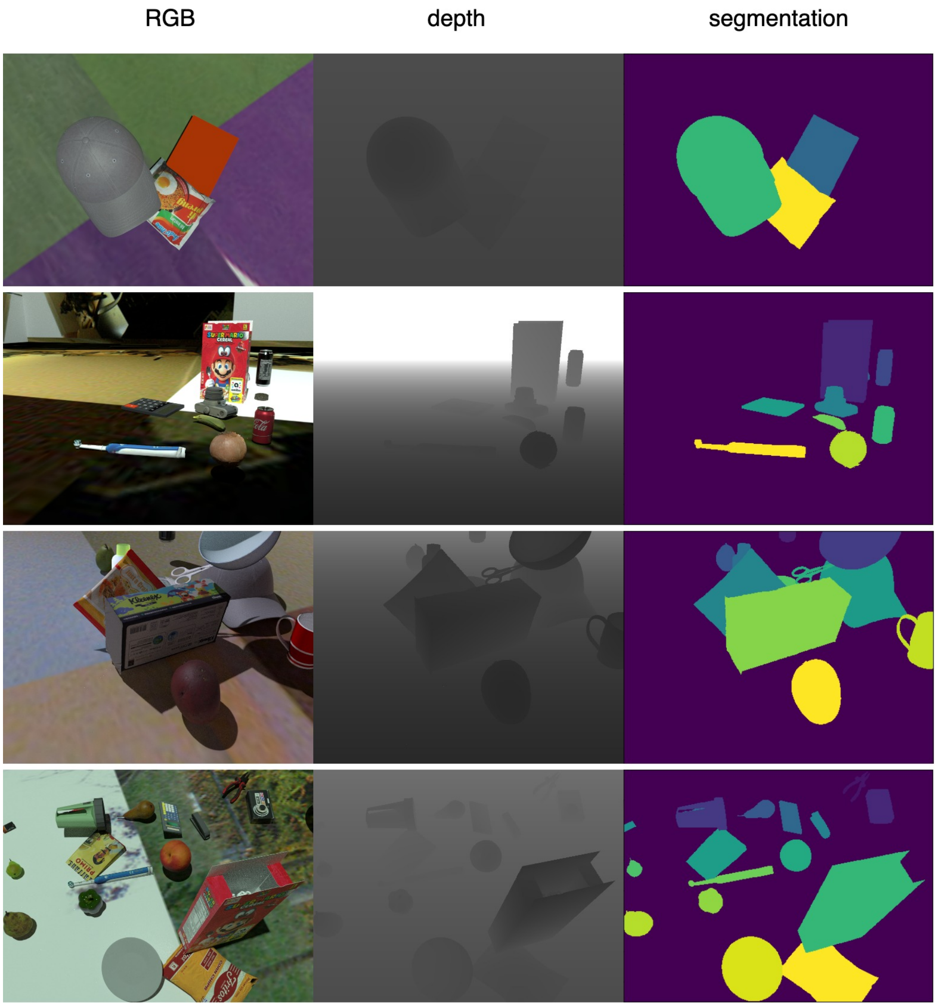


Fig. 3.9 Examples of rendered scenes from synROD dataset. For each, we showcase the RGB, raw depth and segmentation mask image, with increasing level of clutter from top to bottom.

3D model repositories, such as 3D Warehouse and Sketchfab, and retained only models that present texture information to be able to render the RGB modality in addition to the depth. We processed all models to normalize their scale and canonical pose prior to the rendering phase. The final result of the selection stage is a set of 303 textured 3D models from the 51 object categories of ROD, for an average of about 6 models per category.

Rendering 2.5D scenes We render 2.5D scenes using a ray-tracing engine in Blender to simulate photorealistic lighting. Each scene consists of a rendered view of a randomly selected subset of the models placed on a 1.2 × 1.2 meter virtual plane. The camera and light source locations are sampled from an upper hemisphere to ensure a uniform distribution of viewpoints. To achieve this, we followed the recursive division of an icosahedron technique proposed in [327]. The camera and light source distances to the center of the plane varied from 0.7 to 1.5 meters and 2 to 5 meters, respectively. To attain natural and realistic object poses, we dropped each model on the virtual plane using the rigid body physics simulator included in Blender, with a convex hull of the model serving as a collision boundary. The number of objects in each scene ranged from five to 20, resulting in varying levels of clutter. To ensure a balanced dataset, we condition the selection of the models to insert in every scene to the number of past appearances. Finally, we randomized the background of the virtual space containing the objects using MS-COCO [328] images. We produced approximately 30,000 RGB-D scenes with semantic annotation at the pixel level, as shown in Figure 3.9.

3.2.2 Method

In this section, we present our method for RGB-D DA. More specifically, we provide a high-level overview of the method before delving into the relative rotation task. Finally, we specify CNN's architecture and training/testing protocol.

Overview In this work, we aim to address the problem of domain adaptation (DA) in the context of object recognition, where we have access to labeled source data and unlabelled target data. Our approach consists of formulating the problem as a multi-task classification, where we train a neural network to perform a main supervised task and a pretext self-supervised task. Specifically, the main task involves using the supervision of the source data to learn to predict object labels, while the pretext task involves predicting the relative rotation between a pair of RGB and depth images that have been independently rotated. Since the ground truth for this simple pretext task can be generated automatically from the data, we can train the network to predict the relative rotation using both source and target data in a self-supervised fashion. Solving the same task simultaneously on both domains. This allows us to learn inter-modal relations on both domains simultaneously, reducing the distribution discrepancy and yielding domain-invariant features. Furthermore, since rotation prediction is shown to generate semantically-relevant features [20], this pretext task is particularly effective in improving the object class prediction on the target data without the need for direct supervision.

Pretext Task Recognizing the correct orientation of rotating images is a simple yet effective pretext task to learn robust visual representations [20, 329, 175]. This self-supervised task consists in rotating a given image by a multiple of 90° and training a CNN to predict the rotation that has been applied. However, predicting the rotation of an individual image is only possible with datasets such as PACS [19] where the pose of the subject is consistent across samples. For example, the giraffe images in PACS always represent the animal in an upright position. For datasets where the object appears in a variety of poses, predicting the image rotation is an ill-posed problem (see Fig. 3.10). To address this issue, we propose a new task for RGB-D data, where we predict the relative rotation between the RGB x^c and depth x^d images. Let us denote with $rot90(x,i), i \in [0,3]$ the function that rotates clockwise a 2D image x by $i * 90°$. Given an RGB-D sample (x^c, x^d), we select $j, k \in [0,3]$ at random to compute $\tilde{x}^c = rot90(x^c, j)$ and $\tilde{x}^d = rot90(x^d, k)$, and indicate with z the one-hot encoded label indicating the relative rotation between them. More precisely, the relative rotation label is computed as $z = one\_hot((k-j) \ mod \ 4)$, where *one_hot(.)* is the function that generates the one-hot encoding and *mod* is the modulo operator. The pretext task consists of predicting z given $(\tilde{x}^c, \tilde{x}^d)$, or in other words: “how many times should the RGB image be rotated by 90° clockwise to align with the depth image?". Figure 3.11 depicts all of the possible combinations for which a pair of RGB and depth images can be rotated, as well as their relative rotation.

Network architecture Fig. 3.8 shows the structure of the CNN we use for our method.The architecture comprises a feature extractor denoted as E. This module generates RGB-D features, which are then fed as input to both the main head (M) and the pretext head (P). Each of these modules is a neural network defined with differentiable operations, allowing the entire network to be trained end-to-end using standard backpropagation. This design ensures efficient feature extraction and enables effective learning of both the main and pretext tasks simultaneously.

Feature extractor: Following the literature of RGB-D object recognition [72, 18], we use a two-stream CNN with a late fusion approach to generate RGB-D features.

Algorithm 1 RGB-D Domain Adaptation

```
Input:
  Labeled source dataset $S = \{((x_i^{sc}, x_i^{sd}), y_i^s)\}_{i=1}^{N_s}$
  Unlabeled target dataset $T = \{((x_i^{tc}, x_i^{td})\}_{i=1}^{N_t}$
Output:
  Object class prediction for the target data $\{\hat{y}_i^t\}_{i=1}^{N_t}$

  procedure TRAINING(S,T)
      Get transformed set $\widetilde{S} = \{((\tilde{x}_i^{sc}, \tilde{x}_i^{sd}), z_i^s)\}_{i=1}^{\widetilde{N}_s}$
      Get transformed set $\widetilde{T} = \{((\tilde{x}_i^{tc}, \tilde{x}_i^{td}), z_i^t)\}_{i=1}^{\widetilde{N}_t}$
      for each iteration do
          Load mini-batch from S
          Compute main loss $\mathcal{L}_m$
          Load mini-batches from $\widetilde{S}$ and $\widetilde{T}$
          Compute pretext loss $\mathcal{L}_p$
          Update weights of M from $\nabla\mathcal{L}_m$
          Update weights of P from $\nabla\mathcal{L}_p$
          Update weights of E from $\nabla\mathcal{L}_m$ and $\nabla\mathcal{L}_p$
  procedure TEST(T)
      for each $(x_i^{tc}, x_i^{td})$ in $T$ do
          Compute $\hat{y}_i^t = M(E(x_i^{ct}, x_i^{dt}))$
```

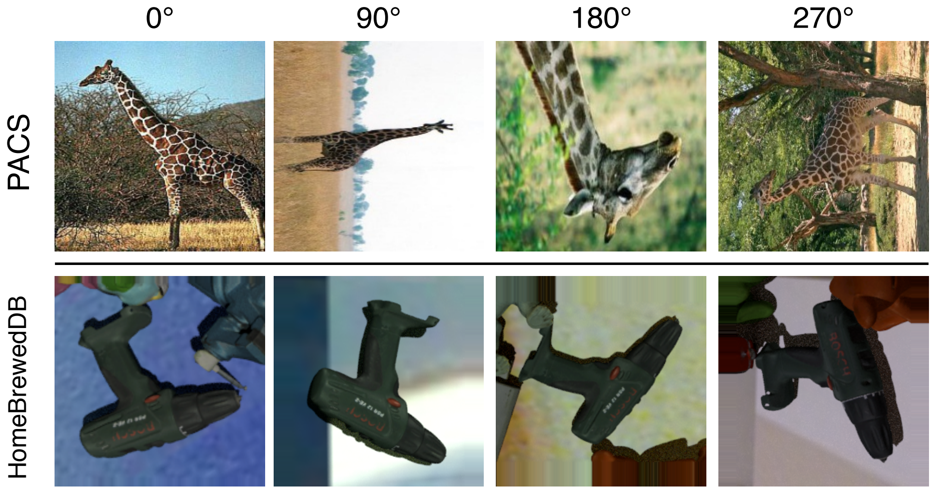


Fig. 3.10 Examples images from PACS [19] (top row) and HomebrewedDB [7] (bottom row) that are rotated by 0°, 90°, 180°, and 270°. While it is possible to infer the rotation of the PACS samples based on the contextual background and prior knowledge of the subject matter, the same cannot be said for the HomebrewedDB samples. This observation highlights the inadequacy of predicting image rotation by analyzing each image in isolation, as proposed in [20]. Hence, this approach is deemed ill-posed.

Specifically, we adopt two identical CNNs, denoted as E^c and E^d, to process the RGB and depth images, respectively. The outputs of these two networks are then concatenated along the channel dimension to compose the final RGB-D feature. For our experiments, We opted for the ResNet-18 architecture for E^c and E^d, without the final fully connected and global average pooling layers. This choice was based on the successful use of ResNet architectures in various computer vision tasks and the availability of pre-trained models for these architectures.

Main head: The network M solves a $\mathcal{C}$-way classification problem, where $\mathcal{C}$ indicates the number of object classes we want to predict. M is composed of three layers: global average pooling (*gap*), a fully connected layer with 1000 neurons (*fc(1000)*), and a fully connected layer with $\mathcal{C}$ neurons (*fc*($\mathcal{C}$)). The *fc(1000)* layer employs batch normalization and ReLU activation, while the *fc*($\mathcal{C}$) layer employs softmax activation.

Pretext head: The network P is responsible for solving the 4-way classification problem of predicting the rotation between the RGB and depth image. It is defined as *[conv(1 × 1,100), conv(3 × 3, 100), fc(100), fc(4)]*, where *conv(k × k, n)* indicates a 2D convolutional layer with kernel size $k \times k$ and n neurons. All convolutional and fully connected layers employ batch normalization and ReLU activation, except for

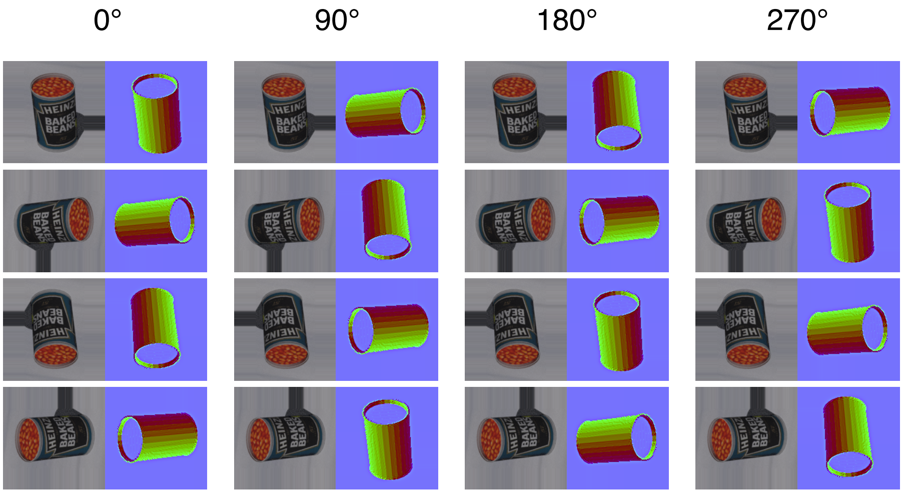


Fig. 3.11 All the possible combinations of RGB and depth rotation for a given relative rotation $\{0°, 90°, 180°, 270°\}$.

the *fc(4)* layer that uses softmax activation. It is worth noting that, unlike M, we use convolutional layers in P to better preserve spatial information. In section 3.2.3, we demonstrate that this approach yields superior performance compared to adopting the architecture of M for both heads.

Optimization We define with $S = \{((x_i^{sc}, x_i^{sd}), y_i^s)\}_{i=1}^{N_s}$ the set of labeled source data and $T = \{((x_i^{tc}, x_i^{td})\}_{i=1}^{N_t}$ the set of unlabeled target data. (x^{*c}, x^{*d}) denotes the pair of RGB and depth images of a sample and y^s denotes the one-hot encoded object class label. From S and T, we can generate a transformed set of source and target data, denoted as $\widetilde{S} = \{((\tilde{x}_i^{sc}, \tilde{x}_i^{sd}), z_i^s)\}_{i=1}^{\widetilde{N}_s}$ and $\widetilde{T} = \{((\tilde{x}_i^{tc}, \tilde{x}_i^{td}), z_i^t)\}_{i=1}^{\widetilde{N}_t}$, that is used to define the relative rotation task. The objective function of our model is given by $\mathcal{L} = \mathcal{L}_m(y^s, \hat{y}^s) + \lambda_p \mathcal{L}_p(z^s, \hat{z}^s, z^t, \hat{z}^t)$, where $\mathcal{L}_m$ and $\mathcal{L}_p$ are respectively the cross-entropy loss of the main and pretext task. λ_p is a hyperparameter that controls the contribution of the relative rotation pretext loss term to the overall loss.

$$\mathcal{L}_m = -\frac{1}{N_s} \sum_{i=1}^{N_s} y_i^s \cdot \log(\hat{y}_i^s), \tag{3.7}$$

$$\mathcal{L}_p = -\frac{1}{\widetilde{N}_s} \sum_{i=1}^{\widetilde{N}_s} z_i^s \cdot \log(\hat{z}_i^s) - \frac{1}{\widetilde{N}_t} \sum_{j=1}^{\widetilde{N}_t} z_j^t \cdot \log(\hat{z}_j^t), \tag{3.8}$$

where $\hat{y}^s = M(E(x^{sc}, x^{sd}))$ and $\hat{z}^* = P(E(\tilde{x}^{*c}, \tilde{x}^{*d}))$. At test time, the predictions of the target data are computed as $\hat{y}^t = M(E(x^{ct}, x^{dt}))$, discarding the pretext head P. The pseudo-code of the algorithm for this process is presented in Algorithm 1.

3.2.3 Experiments

This section covers the experimental protocol and evaluation results of our method. Specifically, it includes the baseline methods that we used for comparison with our method, the implementation details for CNN training, and the presentation of both quantitative and qualitative outcomes obtained from the RGB-D DA analysis.

Baseline methods For our benchmark, we consider four different DA methods: *MMD* [149], *GRL* [154], *Rotation* [175] and *AFN* [151]. The first two are arguably the most widely used and well-established DA methods; *AFN* is chosen as the current state of the art, while *Rotation* is the most relevant to our method.

MMD: The method proposes to minimize the empirical maximum mean discrepancy, a metric that measures the discrepancy between two domain distributions, encouraging the final layers of a neural network to generate domain-invariant features.

GRL: The idea of GRL is to encourage the feature extractor to generate domain-invariant features using adversarial learning. This objective is achieved by jointly optimizing the label predictor and a domain classifier responsible for predicting whether a sample comes from the source or the target domain [6]. Training is performed with the aim of fooling the domain classifier, maximizing its loss through a gradient reversal layer.

AFN: Xu *et al.* [151] pointed out that the main reason behind a difficult classification in the target domain is due to target vectors having smaller feature norms if compared to to that of the source domain. To tackle this issue, the authors proposed to iteratively increase the feature norm of the one-to-last layer of the network for both domains to achieve adaptation.

Rotation: Xu *et al.* [175] proposed a self-supervised task based on geometric image transformations, encouraging the feature extractor to generate domain-invariant features by predicting the absolute image rotation [20].

Since the aforementioned methods are not originally designed for multi-modal data, we use two strategies to evaluate their performance on RGB-D DA. We first adapt each modality individually until convergence, then we freeze the feature extractors and train a fully connected layer on the concatenation of the adapted features (RGB-D). Second, similarly to our method, we apply them to the concatenation of the RGB and depth features generated by the feature extractor E and train the network in an end-to-end fashion (RGB-D e2e). Finally, we report the results for single modalities to see if using multi-modal data is beneficial.

Implementation details The CNN is trained using SGD optimizer with momentum 0.9, batch size 64, learning rate 3×10^{-4}. Following [175, 151], we include entropy-minimization with weight 0.1 as a DA-specific regularization, in addition to the more general weight decay 0.05 and dropout 0.5. The two ResNet-18 weights, E^c and E^d, are initialized with values obtained by pre-training the networks on ImageNet [330], while the rest of the network is initialized with Xavier initialization. All the parameters of the network, including the pre-trained parameters, are updated during training. The input to the network is synchronized RGB and depth images pre-processed according the procedure in [18], where the depth information is colorized with surface normal encoding. This technique prevails as the best non-learned depth colorization method to effectively exploit networks pre-trained on ImageNet and is widely adopted by state-of-the-art methods for RGB-D object recognition [61].

Results Table 3.5 and 3.6 present the quantitative results of RGB-D DA on the two benchmark datasets, synROD→ROD and synHB→realHB. Additionally, Fig. 3.12 provides qualitative insights into the functioning of our method. This empirical evaluation enables us to answer significant research questions.

Are standard DA methods effective on multi-modal data? The results presented in Table 3.5 demonstrate that applying a standard DA method on RGB-D data is not always effective. Specifically, the results reveal that MMD and AFN perform worse when applied on the concatenation of RGB and depth features (“RGB-D, RGB-D e2e) than when applied on the RGB features alone on synROD→ROD. This inferior performance is attributed to the fact that the depth modality is far less informative than the RGB for object recognition when compared in isolation. Consequently, in the absence of an effective strategy to exploit both modalities, the RGB-D case

RGB-D Domain Adaptation			
Method		synROD→ROD	synHB→realHB
Source only	RGB	52.13	51.17
	depth	7.56	15.50
	RGB-D	50.57	49.71
	RGB-D e2e	47.70	49.45
GRL [154]	RGB	57.12	74.74
	depth	26.11	29.52
	RGB-D	59.09	75.23
	RGB-D e2e	59.51	74.95
MMD [149]	RGB	63.68	74.95
	depth	29.34	28.24
	RGB-D	62.10	77.96
	RGB-D e2e	62.57	77.26
Rotation [175]	RGB	63.21	84.46
	depth	6.70	5.62
	RGB-D	63.33	83.99
	RGB-D e2e	57.89	84.15
AFN [151]	RGB	64.63	84.04
	depth	30.72	31.67
	RGB-D	61.19	83.06
	RGB-D e2e	62.40	86.49
Ours		**66.68**	**87.28**
Ours+GRL		**75.11**	**87.81**

Table 3.5 Target Top-1 accuracy (%) of several methods for RGB-D domain adaptation on two datasets with synthetic-to-real shifts, synROD→ROD and synHB→realHB. The highest results are highlighted in **bold**.

can provide lower accuracy than the RGB alone. By comparing the two strategies for applying the baseline methods on multi-modal data (“RGB-D" and “RGB-D e2e"), we observe that no strategy clearly outperforms the other, and the results vary depending on the method and the dataset used. It is also interesting to notice that AFN is not the best performing baseline on RGB-D data, despite being the considered the current state-of-the-art in DA.

Ablation Study			
Method	synROD→ROD	synHB→realHB	avg. drop
Target rotation	63.60	86.32	2.03
FC classifier	64.20	86.49	1.64
Ours	**66.68**	**87.28**	-

Table 3.6 Target Top-1 accuracy (%) of variations of our method for RGB-D domain adaptation on two synthetic-to-real shifts, synROD→ROD and synHB→realHB. The highest results are highlighted in **bold.**

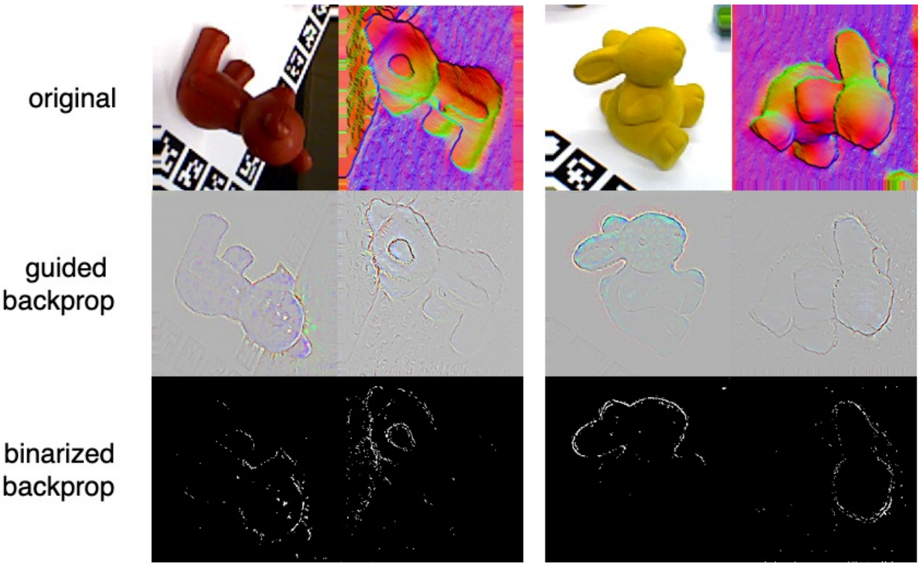


Fig. 3.12 The figure presents a visualization of the significant pixels for predicting the relative rotation. The input of the network is an RGB-D image, labeled as “original." The saliency map of the input is generated using the last layer of the feature extractors E^c and E^d, denoted as “guided backprop" [21]. The “binary backprop" is a binarized version of the “guided backprop" map that highlights peak values in white to aid in visualization. The depth image is utilized with surface normal colorization, following the technique proposed by [18].

Is the relative rotation an effective pretext task to perform RGB-D DA? The results presented in Table 3.5 demonstrate that this method is an effective DA strategy, significantly improving over the "Source only" method. Additionally, the t-SNE [22] visualization of the features of the main head M in Fig. 3.13 confirms

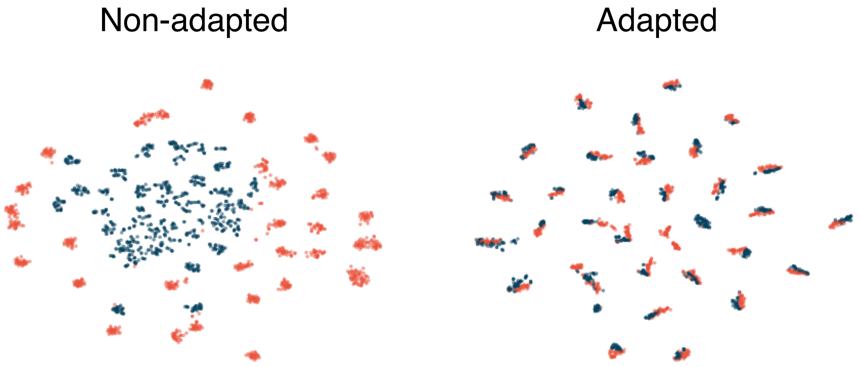


Fig. 3.13 t-SNE [22] visualization of the HomebrewedDB [7]. Features extracted from the last hidden layer of the main head M are used for the plot. Red dots: source samples; blue dots: target samples. When adapting the two domains with our method (right), the two distributions align much better compared to the non-adapted case (left).

that our method aligns the target and source distributions effectively. Moreover, our method outperforms all considered baselines on both datasets, with an improvement of +3.35% and +2.82% on synROD→ROD and synHB→realHB, respectively, compared to the most related method, *Rotation*. This improvement is a consequence of the design of our self-supervised task, which resolves the ill-posedness of Rotation, as discussed in Section 3.2.2.

Is the relative rotation task complementary to existing DA strategies? To assess the complementarity of our method to existing DA strategies, we conducted additional experiments by combining our method with the *GRL* method. To achieve this, we added an extra head for domain discrimination and a domain adversarial loss, as described in [154]. The results of this experiment are presented in Table 3.5 (Ours+GRL). The addition of GRL to our method significantly improved the results on both domain pairs, demonstrating that the two methods are complementary and can be used together to improve domain alignment. The reason behind this improvement is that GRL effectively reduces the domain discrepancy but at the cost of breaking the discriminative structures of the original representations. Our pretext task, on the other hand, helps preserve this structure, as evidenced by the t-SNE visualization in Fig. 3.13. Notably, the improvement was much more significant in synROD→ROD (by +8.43%) than in synHB→realHB (by +0.53%), primarily due to the larger domain gap in synROD→ROD, which provides more opportunities for improvement.

How are the different components of our method affecting the final performance? To understand the influence of different components of our method on the overall performance, we conducted an ablation study. Specifically, we investigated the impact of using both domains to solve the pretext task versus using only the target domain, as well as the effect of defining the pretext head P with the same architecture as the main head M. As shown in Table 3.6, our method achieves significant improvement over the "Source only" baseline when using both domains to solve the pretext task. Predicting the relative rotation of target samples alone also provides some improvement, but not as effective as using both domains, which is due to the higher diversity in the data when using both domains. Moreover, when defining the pretext head P with the same architecture as M, we observed an average drop of -1.64% in accuracy, indicating that using convolutional layers instead of a pooling layer is beneficial to retain the spatial information necessary to predict the relative rotation.

What does the network learn to solve the relative rotation task? In Fig.3.12, we visualize the most relevant pixels for predicting the relative rotation of two example samples in the realHB domain using guided backpropagation[21]. Specifically, we identify which pixels of the RGB and depth input image activate the last layer of E^c and E^d to produce the correct prediction for the pretext task. The results reveal that the most relevant pixels belong to the object itself, rather than the background or other elements in the image. This observation suggests that the network relies on the appearance of the object to make the prediction, rather than learning "trivial" shortcuts [331]. Additionally, the network focuses on the same part of the object, such as the head of the bunny, in both modalities. This finding indicates that the prediction of the relative rotation is based on matching corresponding parts of the object in both RGB and depth modalities..

3.2.4 Conclusion

The present section introduces a novel method tailored to address the challenging task of RGB-D DA. Our proposed approach involves training a network to solve a self-supervised task that predicts the relative rotation between RGB and depth images, in addition to the main object recognition task. To assess the effectiveness of our method, we define two synthetic-to-real benchmarks for object categorization and instance recognition, leveraging the HB dataset and a newly collected dataset called synROD. Our experimental results demonstrate that our self-supervised task effectively mitigates domain shift and outperforms all considered baselines. These results indicate that exploiting inter-modal relations is crucial for achieving successful DA on RGB-D data.

Future research could extend the investigation of the potential of our approach to higher-level tasks, such as egocentric action recognition. This challenging task can greatly benefit from a multi-modal unsupervised domain adaptation technique, particularly by exploiting depth data, which can provide intrinsic attention information to the model. This aspect is particularly relevant as the objects involved in egocentric actions are typically located in close proximity to the user, and therefore their relative position and distance can offer valuable cues for action recognition. Furthermore, a promising direction for future research would be to explore the generalizability of our approach to other visual modalities and foster further research into the demanding problem of domain adaptation for not only RGB-D but also for other multi-modal data.

3.3 From Pixels to Events – DA4E

The last section of this chapter covers the introduction of a new modality i.e. event data, in particular aims to resolve the problem related to the limited availability of this data. To overcome this challenge, researchers often rely on simulated event data as a workaround. However, the use of simulated data raises an open research question regarding its ability to generalize on real data. To answer this, we propose to exploit, in the event-based context, recent Domain Adaptation (DA) advances in traditional computer vision, showing that DA techniques applied to event data helps reduce the sim-to-real gap. To this purpose, we propose a novel architecture, which we call Multi-View DA4E (MV-DA4E), that better exploits the peculiarities of frame-based event representations while also promoting domain invariant characteristics in features. Through extensive experiments, we prove the effectiveness of DA methods and MV-DA4E on N-Caltech101. Moreover, with the motivation to analyze the performance of event cameras in practical applications, we extend our earlier analysis on the RGB-D Object Dataset (ROD) to include the event modality (RGB-E). This allows us to compare the performance of event cameras with traditional RGB-D cameras, providing valuable insights into the strengths and limitations of event-based sensing in practical applications.

Novel bio-inspired devices like Dynamic Vision Sensors (DVS) represent a new category of cameras that operate in a vastly different manner compared to conventional cameras. Rather than capturing images at a fixed rate, event-based cameras function through the *asynchronous* emission of an *event* by each pixel when it detects a local change in brightness. As already described in 2.3, this innovative approach enables these cameras to achieve exceptional levels of performance in terms of high dynamic range, high temporal resolution, and low latency, all while consuming minimal power. In recent years, novel learning approaches utilizing standard computer vision algorithms on event data have demonstrated competitive results compared to traditional approaches [289, 269]. However, training these off-the-shelf deep learning algorithms requires a significant amount of data, which is still limited due to the novelty and high cost of neuromorphic cameras. To overcome the scarcity of data, event camera simulators [44] have emerged as a viable alternative to generate reliable simulated event data. Despite their potential, a key research question arises from this approach: *How well do simulated data generalize to real data?*

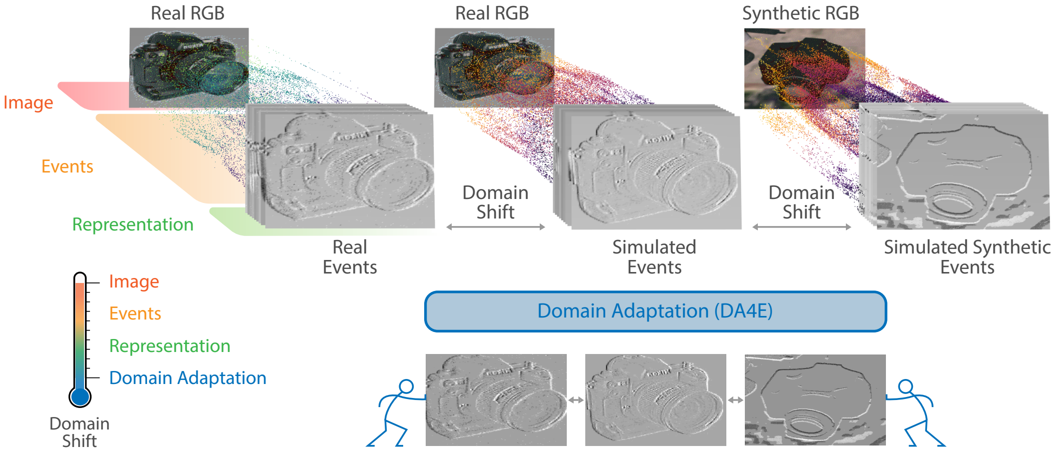


Fig. 3.14 *How can we bridge the Sim-to-Real gap in event-based cameras?* DA4Events leverages unsupervised domain adaptation techniques at the feature level to mitigate this issue. *How else simulated events can be used?* Our proposed approach involves utilizing event data in real-world scenarios, thereby enhancing network robustness through the complementary integration of RGB data.

Researchers have recently made some progress in addressing this issue, as demonstrated in [294] and [295]. In the latter work, the authors proposed reducing the *sim-to-real* gap by manipulating *simulator parameters* during the data simulation phase. Specifically, they acted on the input level to adjust the parameters in order to make the simulated data more realistic and better aligned with real-world data. Our insight is that reducing the sim-to-real gap in neuromorphic vision by operating *at feature level*, during training, leads to more transferable representations, enhancing the generalization performance of deep networks. With this focus, we propose to leverage Unsupervised Domain Adaptation (UDA) techniques [332, 149, 3, 333, 334], as they are specifically designed to align the distribution of features extracted from the source (simulated) and target (real) domains.

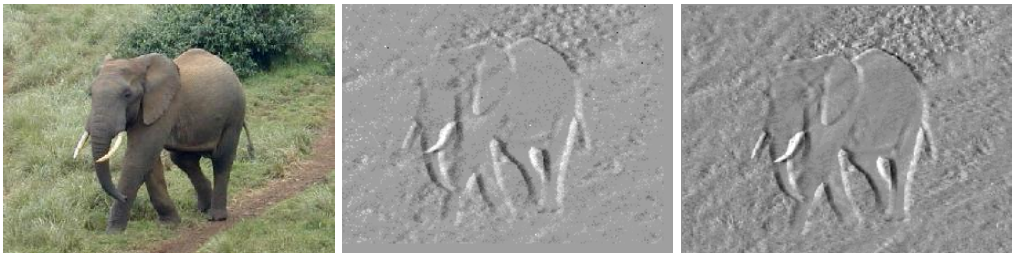
Extensive experimentation on the object classification task using N-Caltech101 [23] and its simulated version Sim-N-Caltech101 [294] validates the efficacy of our proposed method. Our results demonstrate that UDA techniques can effectively bridge the gap between simulated and real event domains, achieving performance on par with a model trained on real-world data. We believe this is a significant step in unlocking event-cameras potential to new tasks, especially those requiring fine-grained annotations, as it enables to exploit the ease of simulation as well as real event sequences that are easy to gather when unlabeled. Our findings unlock novel potential uses of event-based modality, even in situations where it is infeasible to collect it or when working datasets do not provide it. Thanks to the effectiveness of UDA techniques on event-based data, we claim that RGB datasets can be augmented with a *simulated* event modality without compromising performance due to sim-to-real domain shift (Figure 3.14). As mentioned in the previous section, this idea holds particular significance in real-world scenarios, particularly in robotics applications where simulated data is often necessary to compensate the lack of large-scale databases. To demonstrate the quality of simulated events extracted from RGB images and their effectiveness in real-world scenarios, we concentrate on the widely-used RGB-D Object Dataset (ROD)[1], which includes RGB and depth modalities obtained using real sensors, alongside its synthetic counterpart, SynROD[3], generated through digital rendering. We enrich both datasets with simulated events and show that the event modality, even when simulated, produces remarkable results when used in conjunction with RGB data.

In summary, our contributions are the following:

- We propose to bridge the *sim-to-real* gap for event cameras by leveraging UDA techniques, which are currently underexplored in the event-based domain, thereby reducing the problem to a domain shift issue;
- We demonstrate how the domain shift affects different event representations in varying ways and to what extent different UDA techniques can mitigate these issues;
- We propose to deal with event data through a multi-view approach, called *MV-DA4E*;
- We extend the widely-used robotic dataset ROD (along with its synthetic counterpart) by incorporating event data as a new modality, introducing a new RGB-E benchmark for object classification.

3.3.1 DA4Event

As noted by Gehrig et al. [294] and Stoffregen et al. [295], the discrepancies between simulated and real events, as depicted in Figure 3.15, result in decreased performance across various applications, regardless of their representation. This phenomenon is



(a) RGB image (b) Real events (c) Simulated events

Fig. 3.15 Real and simulated events (voxel grid [8]) on a Caltech101 sample.

commonly referred to as the *Sim-to-Real* gap in events. While Gehrig et al. [294] and Stoffregen et al. [295] suggest tackling the issue by addressing event generation, we propose viewing it as a problem of *domain shift*. In this case, the domain shift is not in the visual appearance, as in the well-known *Synth-to-Real* shift existing between rendered RGB images [335] and real RGB ones. Indeed, the primary gap arises from variations in event distribution that correspond to local changes in brightness. Simulators do not account for certain non-idealities that are characteristic of real cameras, such as the minimum threshold C required to trigger an event, or the refractory period of event pixels, which can vary among event cameras.

In this section, we first show that by aligning the feature distribution of the simulated source domain and a target real one, UDA methods effectively reduce the *Sim-to-Real* gap for event cameras, enabling neural networks to take advantage of both simulated data and real unlabeled events during training. Second, we extend our analysis to the *Synth-to-Real* gap, combining both synthetic rendered images and real ones with the corresponding simulated events, showing how the simulated event modality is affected by this shift and how it can benefit from UDA techniques. Additionally, we conduct a third analysis, described in the work [305], that studies the combined effect of *Sim-to-Real* and *Synth-to-Real* shifts. Where we introduce an ad-hoc dataset to study this shift by providing real event data recorded through an event camera rather than obtained through simulation. However, we decided not to include this analysis in this thesis, as it is highly specific to the field of event cameras and strays from the primary focus of this study, which is on multi-modal learning.

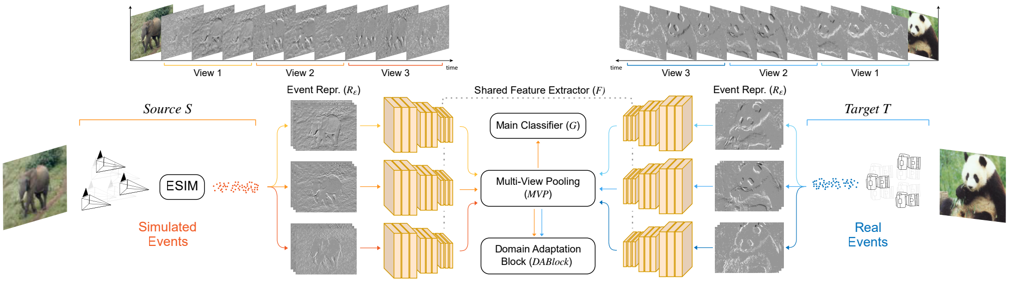

View 1
View 2
View 3
time
time
View 3
View 2
View 1
Event Repr. (R_ε)
Shared Feature Extractor (F)
Event Repr. (R_ε)
Source S
Target T
Main Classifier (G)
ESIM
Multi-View Pooling (MVP)
Domain Adaptation Block ($DABlock$)
Simulated Events
Real Events

Fig. 3.16 Overview of the DA4Event pipeline. **Top** shows the process of extracting an event representation, using voxel grids [8] and three views as an example.**Bottom** details the proposed multi-view architecture (MV-DA4E). During training, two unpaired random batches from source and target domains are sampled and processed separately. When the multi-view approach is not used (DA4E), event representations are fed as a single multi-channel tensor to the feature extractor F, and multi-view pooling is removed. Notice that only source (labeled) data are fed to the classifier G, while both target and source data are fed to the DABlock.

3.3.2 MV-DA4Event: a Multi-View Approach

A common approach to deal with event data is to aggregate the event stream $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}_{i=1}^{N}$ describing the spatial-temporal content of the scene over a temporal period T, into a frame-based representation $\mathcal{R}_\mathcal{E} \in \mathbb{R}^{H \times W \times F}$, thus making events easily processable by off-the-shelf convolutional neural networks (CNNs). While standard RGB images encode spatial (static) information only (R, G, B channels), these frame-based representations additionally carry temporal information, resulting in a variable number of temporal channels as the event sequence is typically divided into several intervals (or bins) to retain temporal resolution, as in a video sequence. For instance, in saccadic motion, commonly used to gather event data from still planar images [23], these channels correspond to the camera response to different move directions.

As a result of carrying temporal information, each temporal channel in a frame-based representation represents a distinct observation of the recorded object, highlighting different aspects (features) of the same. A common practice in computer vision, as well as in the event-based field, is to initialize CNNs with weights pre-trained on ImageNet. However, when using a k-channels representation, where $k \neq 3$, the standard approach is to substitute the first convolutional block with a new one and train it from scratch. This practice not only restricts the exploitation of the pre-trained model but may also be detrimental in a cross-domain scenario. In fact, the first layers of the network are known to be the most affected by the domain shift [336], so, training them from scratch may lead the network to specialize on the source domain, poorly generalizing on the target one. In contrast, transferring pre-trained layers enables the network to benefit from robust low-level features.

Motivated by the aforementioned considerations, we propose to follow a *multi-view* approach to leverage the first pre-trained convolutional layer. This approach involves aggregating the multi-channel event representation into three-channels images, or *views*, resulting in a representation $\widetilde{\mathcal{R}}_{\mathcal{E}} \in \mathbb{R}^{H \times W \times \lceil F/3 \rceil \times 3}$. A multi-view network (Fig. 3.16) has been specifically designed, in which each *view* is fed, separately, to a feature extractor F. The features obtained from each *view* are then combined using a late-fusion approach within a *MVP* module that performs average pooling, producing a $\mathbb{R}^{F_{out}}$ feature vector that is subsequently used throughout the remaining network layers. We believe that fusing the different views at the final layers of the network, rather than at the earliest ones, is beneficial for better generalization. This is because the initial layers of the network are more domain-specific, whereas the later layers carry more task-specific information

Network architecture The proposed network structure is illustrated in Figure 3.16. Events are first obtained using the ESIM [44] simulator in the source domain and directly acquired from the event-based camera in the target domain. These events are then split into B temporal bins, and a sequence of event representations is extracted to obtain a multi-channel volume $\mathcal{R}_{\mathcal{E}}$ with a multiple of 3 channels. The representations are then grouped into views, i.e., 3-channels frames that are treated as images and processed in parallel through a shared ResNet feature extractor F. The set of output features is then combined in the *MVP* module, which performs average pooling both spatially and across the views within features from the same domain, resulting in two feature vectors, one for each domain. Finally, the features from the source domain are used in G for the final prediction and also in the *DABlock*, together with the target features, to perform domain adaptation. It is worth mentioning that during training, two completely random batches of source and target samples are selected with no matching constraints between them.

UDA Algorithms The UDA techniques that we have selected for our analysis encompass a range of different approaches, including adversarial (GRL [332]), discriminative-based (MMD [149] and AFN [333]), and self-supervised methods [337]. For a comprehensive technical explanation of each approach, we refer the reader to Section 3.2.3 (Baseline methods). Additionally, we include a technique called Entropy Minimization (ENT) [334], which is based solely on minimizing classifier uncertainty in the target domain. This involves using a function as a regularization term for the classification loss to represent the uncertainty in the target domain.

3.3.3 Experiments

In this section, we show experiments on object classification tasks in both single- and multi-modal settings. Our assessment of Unsupervised Domain Adaptation (UDA) focuses on event classification, utilizing the N-Caltech101 dataset [23], and we also conduct experiments on multi-modal UDA using the RGB-D Object Dataset [1]. The section contains descriptions of the event representations used in our analysis, along with details regarding the experimental protocol and evaluation results of our method. We compare our findings to baseline methods, presenting both quantitative and qualitative outcomes derived from the RGB-D DA analysis.

Event-Representations In this work we focus on grid-like event representations. Given a stream of asynchronous events $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, the process of extracting a grid-like representation can be described as the conversion from $\mathcal{E}$ to a volume $\mathcal{R}_{\mathcal{E}} \in \mathbb{R}^{H \times W \times F}$ with features F. The methods used in our work for this conversion are summarized in the following.

Voxel Grids. This representation, also known as event volume [8], discretizes the time domain into a traditional image with a fixed number B of channels (where $F = B$) and inserts events into $\mathcal{R}_{\mathcal{E}}$ using interpolation over time. The resulting voxel grid can be expressed as:

$$\mathcal{R}_{\mathcal{E}}^{\text{vox}}(x, y, b) = \sum_{i=1}^{N} p_i k_b(x - x_i) k_b(y - y_i) k_b(b - t_i^*), \tag{3.9}$$

where b is the channel number, t_i^* are the timestamps scaled into $[0, B-1]$, and $k_b(a) = max(0, 1 - |a|)$.

HATS. The Histograms of Time Surfaces (HATS) [264] are a two channel representation format. HATS are built by first dividing the initial event stream grid into C cells of size $K \times K$ pixels each, and then computing, from the events generated by the cell pixels, a grid of $(2\rho + 1) \times (2\rho + 1)$ time surface histograms $\mathbf{h}_{c,p}$ for each polarity p and each cell c. The normalized $\mathbf{h}_{c,p}$ are finally rearranged according to the position of their originating cell and separated into two channels, one per polarity. The ρ parameter is often such that $2\rho + 1 < K$, thus reducing the initial grid resolution. It should be highlighted that temporal resolution is usually lost since the entire temporal window is condensed into a single frame that does not retain temporal resolution.

EST. The Event Spike Tensor (EST) [269] is a trainable representation that can be used end-to-end. It works in a manner similar to a voxel grid, except that it uses the timestamp as the pixel feature and the kernel function used to weigh event contributions is learned by a multi-layer perceptron network. By grouping events by polarity, a two-channel representation can be extracted from each bin.

MatrixLSTM. A recent learnable representation, MatrixLSTM [281], exploits Long Short-Term Memory (LSTM) [338] networks to learn the event accumulation mechanism. In MatrixLSTM [281] the process is similar to EST, but with some differences. The pixel features are computed using a matrix of LSTM cells with shared parameters. Each cell processes the temporal-ordered sequence of events generated by each pixel, and the final output of the LSTM is used as the pixel feature. The number of features can be customized, and bins are optionally used to extract multiple representations.

Implementation details Our proposed method was implemented using the PyTorch autodiff framework. In the N-Caltech101 experiments, we used a ResNet34 [339] as the feature extractor F, while in the ROD experiments, we used a ResNet18 [339]. Both networks were pretrained on ImageNet. To ensure a fair comparison, we used the same network configurations as in [3] for both the object recognition classifier G and the network used in the pretext rotation task. To evaluate the effectiveness of our proposed multi-view approach, we compared it against a baseline with the same architecture that was pre-trained on ImageNet. However, in the baseline model, event representations were directly fed as a single multi-channel tensor without view grouping. To achieve this, the first convolutional layer was replaced with a new randomly initialized convolutional layer matching the number of input channels, and the multi-view pooling stage was removed. During training, event representations and RGB images going through the main backbone F were preprocessed and augmented following the procedure outlined in [3]. The networks are trained using the stochastic gradient descent (SGD) optimizer with a batch size of 32 for N-Caltech101 and 64 for ROD experiments, a weight decay of 0.003, and the domain adaptation (DA) losses' weights are fine-tuned for each event representation and DA method. We report the accuracy scores for the best configurations only, which are averaged over three runs with different random seeds. Input images are normalized using the same mean and variance used for ImageNet pre-training. However, event representations are kept un-normalized as this provides better performance. For voxel grids and EST representations, we use nine bins, resulting in three and six views, respectively, as the latter produces two channels from each bin. For MatrixLSTM, the number of output channels can be customized, and we set the layer to directly produce three-channel output representations and set the number of bins to three as this configuration performs the best. Note that since HATS only provides two channels and does not split temporal frames into bins by default, we cannot apply the proposed multi-view approach.

Results on Sim → Real. We first assess the effectiveness of the UDA algorithms in reducing the domain-shift under the Sim-to-Real scenario using N-Caltech101. Table 3.7 presents the performance of GRL [332], MMD [149], Rotation [337], AFN [333], and Entropy [334] compared to the baseline Source Only. The Source Only method involves training on labelled source data only (*Sim*), and testing directly on unlabelled target data (*Real*), without performing any adaptation strategy. We also report the upper-bound performance obtained by training on real training data and testing directly on it in a supervised fashion (*Supervised*). We evaluate the effect of UDA strategies on two non-learnable event representations (*VoxelGrid* and *HATS*) and two learnable ones (*EST* and *MatrixLSTM*). For each method, we report both the results obtained with (*MV-DA4E*) and without (*DA4E*) the proposed multi-view approach. Our empirical evaluation helps answer the following research questions.

Are UDA methods useful in reducing Sim-to-Real gap?

It is noteworthy that the UDA methods outperform the baseline Source Only for all event representations in almost all cases, as shown in Table 3.7. The UDA methods surpass the baseline by up to 6% on VoxelGrid, 11% on HATS, 6% on EST, and

N-CALTECH101 (SIM $\Longrightarrow$ REAL)					
Method		Voxel Grid	HATS	EST	Matrix LSTM
Source Only	*baseline*	80.99	58.32	80.08	82.21
	MV-baseline	84.59	-	83.07	84.89
GRL [332]	DA4E	83.08	65.38	83.38	82.94
	MV-DA4E	86.77	-	84.03	85.75
MMD [149]	DA4E	86.37	69.86	83.61	84.04
	MV-DA4E	88.23	-	85.36	**88.05**
Rotation [337]	DA4E	79.13	61.52	80.69	83.57
	MV-DA4E	86.63	-	84.49	85.7
AFN [333]	DA4E	84.49	**69.96**	83.59	85.0
	MV-DA4E	88.3	-	85.92	87.59
Entropy [334]	DA4E	87.0	65.58	85.54	85.97
	MV-DA4E	**89.24**	-	**86.06**	86.09
Supervised	*RealEvent*	88.13	76.45	88.17	87.65
	MV-RealEvent	90.09	-	89.25	90.35

Table 3.7 Target Top-1 accuracy (%) of Unsupervised Domain Adaptation methods on N-Caltech101. The highest results are highlighted in **bold.**

4% on MatrixLSTM. Only in one case, Rotation achieves a similar performance to Source Only, which is for VoxelGrid without the multi-view approach. One possible reason for this could be that Rotation mainly benefits the network by enforcing it to focus on the geometric part of the input by solving the transformation. However, event data already contains geometric information (such as movement direction), so Rotation could potentially be unhelpful in certain situations. In fact, the network could learn to find a trivial solution (shortcut) to solve the pretext task [340], such as analyzing the movement direction over the edges.

It is interesting to note that not all event representations suffer equally from the domain shift. For example, HATS seems to be the most affected by the Sim-to-Real shift, with a performance decrease of up to 16% when testing directly on the target domain (Source Only) instead of the source domain (Supervised). This could be due to the fact that when events are represented using HATS, the temporal resolution is

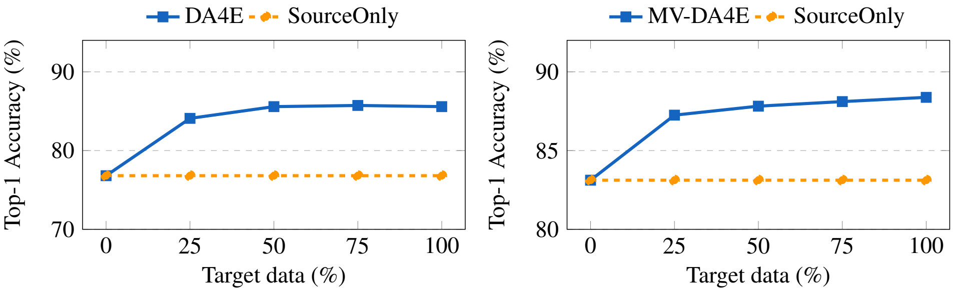
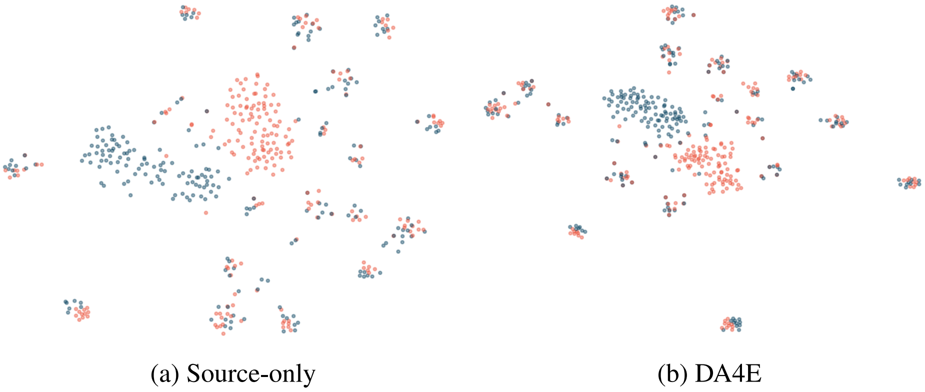


Fig. 3.17 Difference in terms of performance based on percentage (%) of target data used during training, obtained with constant threshold $C = 0.06$.


(a) Source-only

(b) DA4E

Fig. 3.18 t-SNE visualization of N-Caltech [23] features from the last hidden layer of the main classifier. Red dots: source samples; blue dots: target samples. When adapting the two domains with the proposed DA4E (b), the two distributions align much better compared to the non-adapted case (a).

lost. This loss of temporal information may result in a degradation in performance when testing data from a different distribution. Furthermore, in Figure 3.17, we showcase the scalability of our approach when the availability of target data is limited. The figure shows how the performance of the proposed methods changes when only a certain percentage of target data is available during training (25%, 50%, 75%). Interestingly, even when only a small percentage of target samples is available, an improvement of up to 4% over the source only baseline (0% of training target data) is guaranteed. This demonstrates the robustness and effectiveness of the proposed UDA methods in reducing domain-shift in event-based data. Figure 3.18 presents the qualitative results of our approach, including a t-SNE visualization of the source and target samples with and without domain adaptation. Moreover, we utilized the Gradient-weighted Class Activation Mapping (Grad-CAM [24]) to

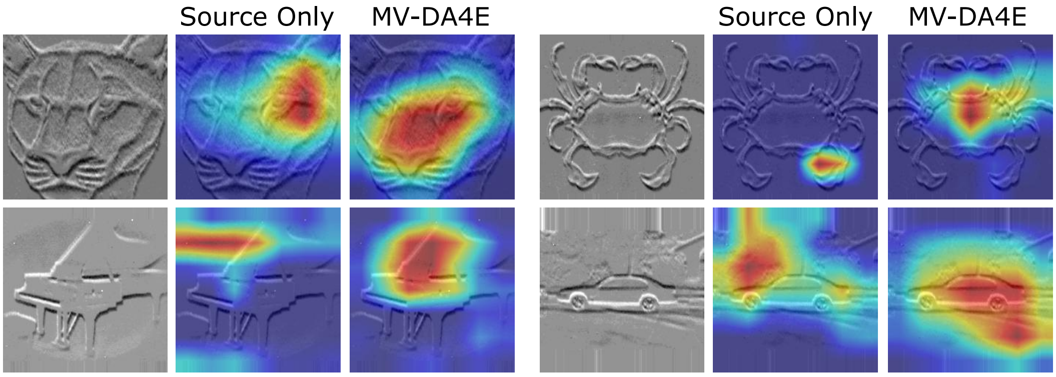


Fig. 3.19 Grad-CAM [24] visualizations on several real N-Caltech101 samples. In each triplet we show the input event representations (voxel grid [8]), the activation maps when the network is trained on simulated data only, and those obtained by training with MV-DA4E.

highlight the regions of the input event representation that the network focuses on for classification. As shown in Figure 3.19, our proposed MV-DA4E approach results in the most discriminative regions for object classification.

Is the proposed multi-view approach MV-DA4E effective?

Table 3.7 displays the substantial performance gains achieved by the multi-view approach *MV-DA4E* compared to the *DA4E* configuration across all experiments, regardless of the DA strategies and representations employed. These results provide compelling evidence supporting the validity of the proposed method, as discussed in Section 3.3.2. What's particularly noteworthy is that *MV-DA4E* not only enhances performance in the cross-domain scenario (Sim-to-Real), but also in the intra-domain (Supervised) setting. This finding suggests that the multi-view approach could serve as a universal tool for handling event representations across various tasks.

How well our approach perform w.r.t. approaches acting on the contrast threshold C?

Some existing methods, such as [294, 295], tackle the *Sim-to-Real* problem by manipulating the threshold value C used by the simulator to generate data. However, since we work with a fixed threshold, one might question whether our approach's success stems from selecting an optimal value for C or from our choice to focus on adaptation at the feature level. To address this question, we conduct experiments using voxel grid as the representation and three different C values: $C = 0.06$ (the starting value used in [294] for domain shift analysis), $C = 0.15$ (estimated fol-

N-CALTECH101				
Baselines		C=0.06	C=0.15 [295]	C$\sim\mathcal{U}$ [294]
Source only	*baseline*	76.81	80.99	82.29
	MV-baseline	83.12	84.59	84.93
Our approach	w/ C values:	C=0.06	C=0.15	C$\sim\mathcal{U}$
GRL [332]	DA4E	80.89	83.08	81.91
	MV-DA4E	84.93	86.77	86.45
MMD [149]	DA4E	83.84	86.37	84.38
	MV-DA4E	86.94	88.23	87.31
ROT [337]	DA4E	80.05	79.13	80.36
	MV-DA4E	86.31	86.63	87.08
AFN [333]	DA4E	84.38	84.49	84.3
	MV-DA4E	87.71	88.3	88.17
Entropy [334]	DA4E	85.26	87.0	85.16
	MV-DA4E	**88.38**	**89.24**	**88.61**

Table 3.8 Target Top-1 accuracy (%) of Unsupervised Domain Adaptation methods w.r.t. to methods that act on the contrast threshold C. The highest results are highlighted in **bold**.

lowing [295]), and $C\sim\mathcal{U}(0.05,0.5)$ (as proposed in [294]). The baselines are the C-only-based methods, where $C=0.15$ reproduces the settings of [295], and $C\sim\mathcal{U}$ reproduces the approach of [294]. Table 3.8 shows that our approach consistently and significantly outperforms the baselines for all C values, demonstrating the effectiveness of addressing DA at the feature level. Moreover, the results indicate that multi-view approaches benefit from UDA techniques in all cases, and even the C-only-based methods benefit from a multi-view approach since it helps to reduce their sensitivity to C variations.

Results on synROD $\rightarrow$ RealROD. As introduced in the previous section 3.2, in the field of robotics, DA techniques are employed to exploit the automatically generated synthetic data that comes with “free" annotations to make accurate predictions on real-world data and overcome the limitations of small-scale datasets. Since the RGB modality primarily encodes texture and appearance information, which are highly affected by domain shifts, adaptation strategies are necessary to mitigate

synROD $\Longrightarrow$ ROD					
Method	RGB	Depth	Event	RGB+E	RGB+D
Source only	52.13	7.56	39.43	52.87	47.7
GRL [332]	57.12	26.11	46.15	55.11	59.51
MMD [149]	63.68	29.34	47.52	62.39	62.57
Rotation [337][3]	63.21	6.70	41.84	66.68	66.68
AFN [333]	64.63	30.72	52.38	66.87	62.4
Entropy [334]	61.53	16.79	49.23	66.23	63.12
Avg	62.03	21.93	47.42	**63.46**	62.86

Table 3.9 Target Top-1 accuracy (%) of the event, RGB and depth modalities, both in single-modal and multi-modal (RGB+E). The highest result is highlighted in **bold**.

this issue. In fact, a recent line of research brought to light that “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness" [341]. With this in mind, we believe that the event modality could be more robust to domain shifts because it encodes additional geometric and temporal information, making it less susceptible to lighting and color variations. Therefore, to verify the effectiveness of using event data extracted from RGB images and their potential in real-world applications, we analyze the performance of the event modality (both single and multi-modal RGB+Event) in the synROD → RealROD scenario (introduced in section 3.2.1). To assess the benefits of the event modality, we compare it to traditional modalities such as RGB and depth.

To conduct our analysis, we selected the VoxelGrid representation and the multi-view approach *MV-DA4E* as they showed superior performance across domains in previous experiments on event data (Table 3.7). Results presented in Table 3.9 clearly demonstrate that the event modality is more robust than the depth modality, and that UDA techniques are effective in this setting. Specifically, we found that the event modality is less sensitive to domain shifts compared to the depth modality, both in single modal (7.56% *vs* 39.43%) and multi-modal RGB+E (47.7% *vs* 52.87%) scenarios, without applying any DA techniques. Interestingly, when combined with the RGB modality, event data slightly improves model accuracy, in contrast to the depth modality, which causes performance degradation as the network struggles to exploit the complementarity of the two modalities. Additionally, we found that UDA performance on RGB-E is generally better than that of RGB-D.

3.3.4 Conclusions

In this work, we propose an alternative way of answering a very recent research problem regarding how to bridge Sim-to-Real gap for event cameras arising from event generation. By seeing the problem under a new perspective, the domain shift, we show that Unsupervised Domain Adaptation (UDA) techniques working at the feature level are an effective way of tackling this issue, w.r.t. previous works that act on the input level. Moreover, we propose a multi-view approach to deal with event representations, which outperforms existing methods and proved to work well in conjunction with other UDA strategies. Finally, with our analysis, we demonstrate the potential of event data in robotics applications, particularly in the challenging Syn-to-Real scenario. We show that event data, even if simulated, can effectively encode geometric information and is less sensitive to domain shifts compared to traditional modalities such as depth. Our experiments also highlight that the combination of event data with RGB can improve model accuracy, enriching standard RGB information with additional geometric information. We validate both approaches through extensive experiments on the N-Caltech101 dataset and the popular RGB-D Object Dataset (ROD).

Our analysis has also uncovered two critical findings that should capture the attention of the research community. Firstly, our results demonstrated that using multiple modalities as input to the network does not always lead to better results, as evidenced by the limited improvement obtained by the RGB-Depth combination. This finding raises questions about the optimal design of multi-modal architectures and the importance of carefully selecting the modalities that complement each other to achieve better performance. Secondly, our study has confirmed the event modality as a new and promising source of information for robotics applications, with its unique ability to encode geometric and temporal information. While we have shown the potential of this modality, further research is necessary to fully exploit its capabilities, particularly in tasks where temporal information plays a central role. These insights provide exciting opportunities for future research into the event modality's role in robotics and other emerging fields.

Chapter 4

Multi-Modal Learning for Egocentric Vision: First Person Action Recognition

This chapter presents a comprehensive investigation into the field of multi-modal learning for first-person action recognition. A significant challenge in this context arises from the necessitating analysis of motion and the extraction of relevant information alongside visual cues to accurately identify actions. While conventional approaches rely on optical flow to separate motion from the visual context and enhance generalization, their applicability in real-world scenarios, particularly for wearable devices, is limited. This chapter primarily investigates two key areas. Firstly, we propose novel techniques to enhance existing models' ability to extract motion information, utilizing advanced self-supervised methods. Secondly, we introduce a new training approach that effectively integrates both visual and non-visual modalities, thereby improving the overall generalization and adaptability of networks for action recognition tasks. Additionally, we explore the potential of event-based cameras as a viable alternative to optical flow for capturing motion information. We introduce a strategy to adapt existing action recognition models to effectively utilize event-based data. By conducting a comprehensive investigation in these areas, this chapter aims to contribute to the advancement of multi-modal learning for egocentric vision, enabling more robust and practical approaches to first-person action recognition.

4.1 Single-stream Multi-Modal Fusion – SparNet

The complex nature of egocentric videos presents several challenges that can impact the performance of the standard action recognition models, especially when relying solely on RGB data. To overcome these limitations, state-of-the-art approaches incorporate optical flow, that helps to capture motion information and improve performance. However, obtaining high-quality optical flow is still a computationally intensive operation, making it unfeasible for use in online applications. In this section, we propose a novel approach for egocentric video understanding that overcomes the limitations of existing two-stream approaches by leveraging a single-stream architecture that jointly learns motion and appearance information. To this end, we introduce a self-supervised block utilizing a pretext motion segmentation task that interweaves motion and appearance knowledge. Unlike previous two-stream methods that require optical flow during training and testing, our approach is able to achieve comparable performance without the need for optical flow information during testing. We evaluate our approach on several publicly available databases and demonstrate its effectiveness in comparison to state-of-the-art methods.

Recent advances in wearable devices have sparked a growing interest in FPAR, enabling the capture of activities while users are in motion, without the need for external sensors. However, transitioning from third-person to first-person action recognition presents unique challenges. Firstly, the presence of strong egomotions introduces complexities as the cameras are mounted on the actor's body. Secondly, the limited availability of pose information in first-person videos hinders the analysis of actions compared to third-person videos. Lastly, occlusion and partial visibility of arm trajectories and hand gestures are common in first-person videos. Thus, it is critical to extract as much information as possible from video frames about the objects being manipulated, their position, and the motion data encoded in the video (since, for instance, the correct interpretation of the actions of "opening" and "closing" a bottle, Fig. 4.1, depends merely on the hand motion direction).

To address the last issue, a popular strategy is to combine two types of information: the visual appearance of the object of interest, which is modeled by the spatial stream that processes RGB images, and motion information, which is handled by the temporal stream, which takes as input the optical flow extracted from adjacent frames. Furthermore, attention modules are frequently integrated into the basic two-stream architecture in order to identify the more informative frames and regions

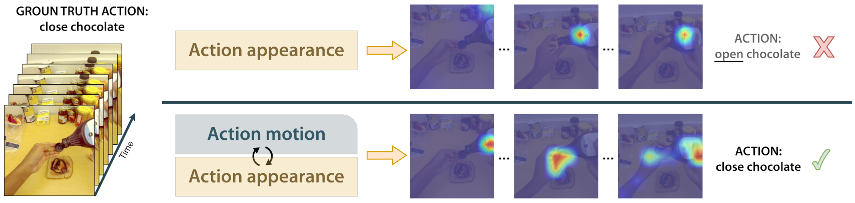


Fig. 4.1 By leveraging a self-supervised motion prediction task during training, at test time, SparNet can jointly exploit motion and appearance information using a single RGB stream. The result is a leaner architecture that better focuses on the relevant elements for action recognition in egocentric vision, as can be seen from the comparison between the Class Activation Maps when the auxiliary task is used (lower path) or not (upper path)

for the task [35, 34]. Despite their effectiveness, these methods have three significant flaws. Firstly, as previously mentioned, they heavily rely on high computational cost modalities like optical flow. Secondly, the spatial and temporal features that capture the appearance and motion information of the object of interest are learned independently, and their final predictions or feature embeddings are fused at the network's end using simple weighted sums or concatenation [342, 34, 35]. This approach is suboptimal because it fails to effectively model the correlated spatial-temporal relationships between the two features. Thirdly, as two-stream approaches strive to improve performance, the overall architecture's parameter count grows significantly, which could be a limiting factor for wearable devices where memory is a scarce resource.

In this section, we address these issues by moving beyond the two-stream paradigm and proposing an architecture that couples the modeling of motion and appearance information within a single RGB stream by leveraging one or more motion-prediction (MP) self-supervised tasks. These tasks "force" the backbone to learning an image embedding that focuses on object movements, a piece of information that is beneficial for the main task of FPAR. Thanks to the use of the auxiliary tasks, this information is directly encoded in the inner layers of the backbone, hence leading to an intertwined learning of appearance and motion features. We demonstrate the effectiveness of this idea not only through our results but also by the results obtained from including these MP pretext tasks in other recent models, such as Ego-RNN [34] and LSTA [35]. Our resulting architecture is relatively straightforward, comprising a standard backbone (i.e., a ResNet-34 in our experiments), followed by a standard ConvLSTM. The auxiliary task heads consist of shallow architectures.

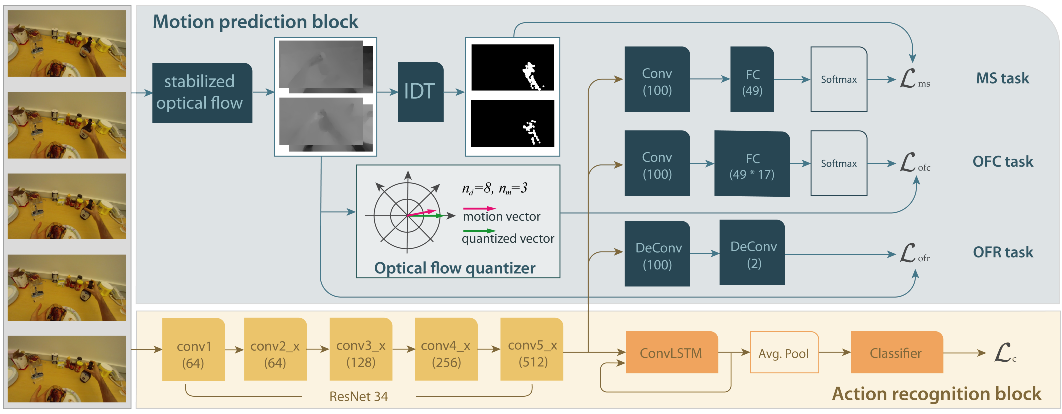


Fig. 4.2 SparNet architecture. The action recognition block computes image embeddings and solves the classification task. The motion prediction block injects into the backbone motion information that results in a richer embedding, jointly encoding motion and appearance information. We compensate for egomotion effects with the use of stabilized dense optical flow and an IDT module [25]. The OF quantizer block discretizes the input motion vectors into a finite number of classes according to the values of the n_d and n_m parameters (respectively, 8 and 3 in this example, for a total of 17 classes).

Due to its simplicity, our architecture can be trained end-to-end in a single stage, in contrast to various other two-stream methods [34, 35]. Moreover, it can utilize a smaller number of frames than what was used in previous works without any adverse effect on performance. We refer to our architecture as Self-supervised first Person Action Recognition network - SparNet.

To summarize, the contributions of this section are as follows: (i) we introduce a novel set of motion prediction self-supervised tasks specifically designed for egocentric action recognition; (ii) we address the problem of how to effectively leverage over a self-supervised branch to jointly encode spatial and motion information by identifying the features that are most suited to solve this task; and (iii) we showcase the effect of each component of SparNet with a quantitative and qualitative ablation study.

4.1.1 Architecture Overview and Details

SparNet: Overview. In the proposed architecture's basic version (as shown in Figure 4.2, action recognition block), we initially extract a limited number of N sparse representative RGB frames from each input video segment. These frames' appearance embeddings, obtained through a standard CNN backbone, are then fed into a ConvLSTM network. The network's output is passed through an average pooling layer and subsequently through a fully connected (FC) layer for classification. Although utilizing a small number of frames helps reduce the computational burden of the model, the resulting features lack crucial motion information necessary for accurate recognition, which is exploited by two-stream approaches that incorporate explicit optical flow data.

To tackle this issue, we introduce a regularization technique for representation learning in FPAR by extending the basic architecture to a multi-task network. This network is required, at train time, to solve jointly two different problems: the action recognition task and a motion-prediction (MP) auxiliary task. We formalize the latter as a self-supervised problem that, given a single (and static) RGB frame as input, tries to answer one (or both) of the following questions: which parts of the image are going to move? And in which direction?

In our approach, we formulate the motion segmentation task (MS), which involves identifying moving parts, as a labeling problem aimed at minimizing discrepancies between a *motion map* that labels pixels as either *moving* or *static*, and the object movements predicted by the network based on a single static RGB frame. These unsupervised motion maps are obtained from the input video segment using a technique inspired by [343] and leveraging Improved Dense Trajectories (IDT) [25] for extracting "stabilized" motion information. The main idea of IDT is to compensate for strong camera motion and shake typically observed in egocentric videos by estimating the homography that relates adjacent frames. Subsequently, keypoints that can be reliably tracked for at least eight frames and are not identified as camera motion are labeled as *moving*.

The objective of the second sub-problem of MP is to estimate the "stabilized" flow, which refers to the dense optical flow computed after compensating for camera motion between consecutive frames, from a static RGB input image. As optical flow is a continuous function, we initially formulate its estimation as a regression problem, referred to as Optical Flow Regression (OFR). Alternatively, we also convert it into a classification problem, termed Optical Flow Classification (OFC). For OFC, we quantize the per-pixel motion vectors as follows: first, we extract the magnitude and direction of each motion vector for every pixel. Then, we discretize the angle into a set of n_d directions uniformly distributed in the interval $[0, 2\pi]$. The magnitude is discretized by clamping it to a maximal value v_{clamp}, normalizing it to one, and dividing the interval $[0,1]$ into n_m values (including the extremes), with the constraint $n_m \geq 2$. Consequently, each pair (d_m, d_d) of discretized magnitude and direction values is assigned a unique label, except for motion vectors with magnitudes close to zero (whose orientations tend to be meaningless), which are assigned to the same class. Hence, the total number of classes is $n_d \cdot (n_m - 1) + 1$, and the self-supervised task involves estimating the correct flow labels.

Both approaches, regression and classification, have their own strengths and weaknesses. Regression has the potential for higher accuracy, but it can be challenging to solve and may result in sub-optimal solutions due to smoothing effects as noted in [344]. On the other hand, classification tends to have more stable convergence, but it introduces quantization errors. One possible solution is to combine the advantages of both methods. Similar benefits can be expected by integrating motion segmentation (MS) with optical flow estimation. MS identifies points that exhibit stable and coherent motion over a longer temporal interval than just two adjacent frames, thereby reducing the impact of noise on dense optical flow. In contrast, optical flow estimation can provide robust identification of moving parts in an image, along with information about the direction of their motion. By coupling MS and optical flow estimation, we can potentially achieve a more robust and accurate motion analysis approach that leverages the strengths of both regression and classification methods.

Regardless of the pretext task chosen for the MP module, whether it's MS, OFR, OFC, or a combination of them, its main objective is to facilitate the learning of motion cues within the appearance stream. We posit that by processing inputs with these characteristics, the ConvLSTM can extract a more meaningful global video representation. This representation encompasses not only appearance information but also short and long-term motion dependencies among frames. This is in contrast to the vanilla appearance embeddings, which may not fully capture the rich motion information present in the video. By leveraging the MP module, we aim to enhance the global video representation and enable the network to better understand the complex interplay between appearance and motion cues, leading to improved performance in video analysis tasks.

SparNet: Details. Let $\mathcal{S}$ be a training set consisting of samples $S_i = \{H_i, y_i\}_{i=1}^{n}$, where H_i is a set of N timestamped images $\{(h_i^k, t_i^k)\}_{k=1}^{N}$ uniformly sampled from the video segment. Let also $x = f_M(H|\theta_f, \theta_c)$ be the embedding of sample S computed by our model M, where parameters θ_f and θ_c define, respectively, the image embedding and the classification spaces. Finally, let $g(x)$ be a class probability estimator on the embedding x.

The action recognition and the MP task share a common trunk that is completed by two task-specific "heads". The first objective of the learning step consists in minimizing the categorical cross-entropy classification loss $\mathcal{L}_c$:

$$\mathcal{L}_c(x,y) = -\sum_{i=1}^{n} y_i \cdot log(g(x_i)) \tag{4.1}$$

Together with the objective mentioned above, we ask the network to solve an MP task, whose head can take different shapes according to the specific sub-problem chosen (or combination of sub-problems) and whose input is always the output of the backbone.

The MS task is characterized by a shallow head composed by a single convolutional block, aimed at both adapting the features to the MS task and reducing their channel number. This head ends with a fully connected layer of size s^2 followed by a softmax, and it is trained with a loss $\mathcal{L}_{ms}$ based on the per-pixel cross entropy between the computed label image l_{ms} and the ground truth m (which is first downsampled to a size $s \times s$ and then vectorized). The estimated motion map l_{ms} is obtained as a function of both image embedding z, which depends only on θ_f, and MS head parameters (θ_{ms}). Thus, the $\mathcal{L}_{ms}$ loss can be defined as:

$$\mathcal{L}_{ms}(z,m) = -\sum_{i=1}^{n}\sum_{k=1}^{N}\sum_{j=1}^{s^2} m_i^k(j) \cdot log(l_{ms,i}^k(j)) \tag{4.2}$$

where m is the ground truth.

The OFR task aims at regressing the two separate horizontal and vertical components of the optical flow. Its head is composed by a stack of two deconvolution layers (and ReLU activation functions) that learn a nonlinear upsampling to a final size of $r \times r \times 2$. The head is trained by minimizing the Mean Squared Error (MSE) between the predicted optical flow (which is a function of θ_f and OFR head parameters, θ_{ofr}) and the ground truth (which is downsamples to a size $r \times r \times 2$). The $\mathcal{L}_{ofr}$ task loss can be defined as:

$$\mathcal{L}_{ofr}(z,of) = \sum_{i=1}^{n}\sum_{k=1}^{N}\frac{1}{r^2}\sum_{j=1}^{r^2}||of_i^k(j) - y_i^k(j)||_2^2 \tag{4.3}$$

where of is the ground truth and y is the estimated optical flow.

Finally, the structure of the OFC head is identical to that of the MS task, with the only exception that its ground truth is obtained by first downscaling the optical flow to a size $s \times s$ and then quantizing it. The OFC loss is then defined as:

$$\mathcal{L}_{ofc}(z,of) = -\sum_{i=1}^{n}\sum_{k=1}^{N}\sum_{j=1}^{s^2} Q(of_i^k)(j) \cdot log(l_{ofc,i}^k(j)) \tag{4.4}$$

where l_{ofc} is the computed label image (depending from θ_f and the OFC head parameters θ_{ofc}) and Q is the one-hot vector containing the result of the quantization of the of ground truth.

The optimal model of SparNet is achieved by simultaneously solving two separate optimization problems: minimizing $\mathcal{L}_c$ (the classification loss) and the loss of the chosen MP problem (MS, OFR, OFC, or a combination), each with an additional L^2 weight regularization term. The architecture is designed such that the weights of the MP head are updated only through the backpropagation of its loss error, while the classification parameters (θ_c) are updated only through the $\mathcal{L}_c$ error. Both losses contribute to updating the weights of the backbone (θ_f) through a weighted combination of their gradients, with the weights being hyperparameters of the method. In case of multiple auxiliary tasks, each task is optimized independently, and their weighted gradients are combined with that of the action classification head to optimize the backbone. The overall architecture is illustrated in Figure 4.2.

SparNet: Implementation. While SparNet network can leverage over many possible backbones, we choose for our experiments a ResNet-34 model pre-trained on ImageNet, which is both a lightweight and powerful backbone. The motion-prediction heads receive in input the features extracted from the `conv5_x` block of the ResNet (whose size is $7 \times 7 \times 512$). The MS and OFC heads reduce the feature channels to 100 and the size s^2 of their resulting ground truth is, therefore, 49. Both OFR deconvolutional blocks have a 3×3 kernel and a stride of 2. The first reduces the input feature channels to 100 and the second to two (i.e., the estimated horizontal and vertical optical flow displacements). The final value of r is 35. As for the ground truths, we compute the dense optical flow with the Gunnar-Farneback method [345].

4.1.2 Experiments

In this section, we provide details about the implementation used in our experiments. We then perform an ablation study to demonstrate the effectiveness of the proposed self-supervised MP tasks and our single-stream approach. We also analyze their effects on different models. Finally, we discuss the results obtained on four standard first-person action recognition datasets (described in Section 2.4.1), which highlight the strength of SparNet in these benchmark datasets.

Implementation details SparNet is trained end-to-end on a single stage. The ConvLSTM cell in our implementation has 512 hidden units for temporal encoding, and it is initialized following the approach used in [34]. During training, we employ different learning rates for different architectural blocks, including the backbone, MS head, ConvLSTM, and the final classification layer. The number of training epochs varies depending on the dataset, with 400 epochs for GTEA-61, 70 epochs for EGTEA+, 100 epochs for FPHA and EK. We use the ADAM optimization algorithm for training, and the batch size is set to 4 for GTEA-61 and 8 for the remaining datasets. We decompose each input video segment into $N = 7$ frames for GTEA-61 and FPHA, and $N = 11$ frames for EGTEA+ and EK. The frames are uniformly sampled in time. We set $v_{clamp} = 15$ and $n_m = 4$, with a varying number of angular subdivisions ($n_d \in 8, 16, 20$). Input images are resized to a height of 256 pixels, while maintaining the aspect ratio to update the width accordingly. The actual training input is a random crop of size 224×224 pixels. Ground truth for MP tasks is computed by first scaling all videos to a fixed height of 540 pixels. During training, we apply data augmentation techniques proposed in [346]. At test time, we feed the network with the central crop of the frames. To ensure result reproducibility, we conduct three runs for each experiment, with the same constant seed across different datasets and parameters. Therefore, unless stated otherwise, the reported results for SparNet are the average accuracy over these three runs.

GTEA-61		EGTEA+		FPHA	
EleAttG [347]	66.77	RULSTM [46]	60.20	H+O [348]	82.43
TSN [349]	69.93	Ego-RNN [34]	60.76	Gram Matrix [350]	85.39
Ma et al. [89]	73.02	LSTA [35]	61.86	ST-TS-HGR-NET [351]	93.22
Ego-RNN [34]	79.00	3DConv MTL [36]	65.70		
LSTA [35]	80.01	Two-stream I3D + STAM [131]	65.97		
Baseline	80.18	Baseline	63.96	Baseline	94.32
SparNet-MS	80.51	SparNet-MS	66.15	SparNet-MS	96.41
SparNet-OFR	80.14	SparNet-OFR	64.22	SparNet-OFR	95.07
SparNet-OFC	81.17	SparNet-OFC	67.36	SparNet-OFC	96.41
SparNet-OFR+OFC	80.51	SparNet-OFR+OFC	**67.52**	SparNet-OFR+OFC	96.35
SparNet-MS+OFC	**81.39**	SparNet-MS+OFC	67.44	SparNet-MS+OFC	**96.70**

Table 4.1 Top-1 accuracy (%) achieved by various state-of-the-art methods on GTEA-61, EGTEA+, and FPHA datasets. The highest results are highlighted in **bold**.

Ablation Study In this section, we comprehensively evaluate SparNet on the first split of the EGTEA+ dataset using as baseline the action recognition block in Figure 4.2. Specifically, we study the following aspects.

EGTEA+	
Method	Accuracy (%)
SparNet-MS (7 frames)	67.05
SparNet-MS (11 frames)	**68.43**
SparNet-MS (16 frames)	67.48
SparNet-MS @ `conv4_x`	66.15
SparNet-MS @ `conv5_x`	**68.43**
SparNet-MS @ `Output ConvLSTM`	67.68

Table 4.2 Top-1 accuracy (%) achieved by our method using a different number of representative frames and input features. The highest results are highlighted in **bold**.

Sparse sampling. We start by analyzing the effect of the number N of input frames used for action recognition (in both train and test). Using the MS task as a reference and varying N in the range of $[5,25]$, we observed that there were no significant differences in performance for values between 9 and 11. However, the error rate started to increase slightly for smaller and larger values of N, as shown in Table 4.2 (only a selection of significant values are reported). This observation is consistent with findings from other MP tasks, which we do not show here for brevity. These results are in line with the findings in [349], suggesting that a dense temporal sampling may result in highly redundant information that is unnecessary for capturing the temporal dynamics of the video, while too few frames can lead to the loss of relevant cues for the current action. Therefore, in our main experiments, we heuristically adapted the value of *N* based on the average segment length of the analyzed dataset.

MP taks: input features. The choice of input features for the auxiliary MP tasks plays a crucial role in their effectiveness. In our experiments, we explored two options: using the output features of the residual blocks of the ResNet backbone or utilizing the spatio-temporal representations obtained from the ConvLSTM. As shown in Table 4.2, the `conv5_x` features significantly outperformed the lower layer features (such as `conv4_x`), indicating that leveraging high-level and more structured information from deeper layers of the backbone is beneficial for the MP tasks. However, the accuracy obtained from the ConvLSTM features was lower than that of the `conv5_x` features. We hypothesize that this may be due to the spatio-temporal processing capability of ConvLSTM, which makes the solution of the MP task easier. As a result, the backbone may incorporate less motion information in its embeddings when using ConvLSTM features, which could have negative effects on the main FPAR task. This suggests that finding the right balance between incorporating motion information and maintaining discriminative features for the main task is crucial in designing effective MP tasks.

Impact of the MP tasks. Table 4.3 presents an ablation study of different variants of SparNet, along with the total number of parameters and GFLOPS of the resulting architecture, as well as state-of-the-art results on the same split. This analysis allows us to understand the individual and mutual contributions of the different MP tasks. From the results, it is evident that both individual and combined MP tasks contribute to improving the baseline performance. However, the extent of improvement varies among the tasks. OFR is the most challenging task, as indicated by its lower contribution to the final FPAR task. On the other hand, the choice of translating the optical flow estimation into a classification problem proves to be effective. Furthermore, the mutual contribution of the individual MP tasks helps in making the MP problem more robust, which, in turn, provides better integration of appearance and motion information in the backbone. It is worth noting that the overall performance of SparNet is competitive with state-of-the-art results on the same split, indicating the effectiveness of the proposed approach in integrating motion information for fine-grained action recognition.

ABLATION STUDY EGTEA+			
Method	Acc (%)	Param (M)	GFLOPS
baseline	65.46	24.34	41.52
SparNet-MS	68.43	24.87	41.55
SparNet-OFR	65.73	24.80	43.01
SparNet-OFC ($n_d = 8$)	69.32	30.39	41.61
SparNet-OFC ($n_d = 16$)	69.49	36.16	41.68
SparNet-OFC ($n_d = 20$)	68.96	39.04	42.21
SparNet-OFR+OFC ($n_d = 16$)	69.57	36.62	43.17
SparNet-MS+OFC ($n_d = 16$)	**69.80**	36.70	41.71
Ego-RNN RGB [34]	-	24.34	94.36
Ego-RNN [34]	-	45.71	98.31
LSTA RGB [35]	57.96	41.22	114.92
LSTA [35]	61.86	62.59	118.86
Two-stream I3D + STAM [131]	68.60	-	-
3DConv MTL [36]	68.99	-	-

Table 4.3 Top-1 accuracy (%) achieved by variations of our method compared with others by various state-of-the-art methods on EGTEA+ (split 1, $N = 11$). The highest result is highlighted in **bold**.

Concerning the computational burden of the MP tasks, we can say that their effect is in general minimal, exception made for OFC that requires a larger number of parameters for the classification (but still a limited increase in terms of GFLOPS). We recall that the baseline numbers are those required at test time (when the MP tasks are disabled). It can be seen that the relative increase in term of parameters (GFLOPS) is 2.2% (0.1%) and 1.9% (3.6%) for, respectively, MS and OFR and up to 60.41% (0.5%) for OFC (when $n_d = 20$). These numbers can be compared by those expressed by Ego-RNN and LSTA, whose GFLOPS are substantially higher (an increase between 127.3% to 186.3%) in both their single and two-stream versions and in both train and test time.

MP tasks and other models. One possible question is if the proposed MP tasks can be beneficial to other models too. To this end, we performed a detailed analysis of their effects on Ego-RNN RGB [34] and LSTA-RGB [35]. Both methods converge to the same baseline of SparNet when the CAM is deactivated (in Ego-RNN RGB) or a vanilla LSTM cell is used instead of the proposed LSTA cell (in LSTA-RGB). For these experiments, we modified both architectures adding various MP tasks, feeding them with the `conv5_x` features and 25 frames in input, as in their original papers.

GTEA-61			
Method	Single-stream (SS)	SS + MS	SS + OFC
Ego-RNN [34]	63.79	68.97	68.10
LSTA [35]	65.80	66.96	67.24

Table 4.4 Top-1 accuracy (%) achieve by applying the motion segmentation (MS) task on Ego-RNN [34] and LSTA [35].

We present results obtained on the second split of GTEA-61. For the sake of brevity, we report the results obtained with MS and OFC. For a fair comparison, single-stream results are those obtained in our experiments, which, despite our efforts, could not replicate those presented in [34] and [35]. We also underline that, since both methods retrain merely the last residual block of the backbone and not the whole ResNet as in our case, the MP effect is not back-propagated to the lower backbone layers, preventing them from supporting the higher ones in learning new features that are more focused on the actual FPAR task. Nonetheless, we think the numbers in Table 4.4 highlight the effect of MP on these models and showing that the effectiveness of our approach is not limited to SparNet.

Experiments on GTEA-61, EGTEA+ and FPHA In our experiments on the GTEA-61 and EGTEA+ datasets, we followed the protocols defined in [34, 35], which require reporting the final average accuracy over different and fixed non-overlapping training and test sets. For the FPHA dataset, we followed the 1:1 protocol, which defines fixed training and test sets. We compared SparNet with several state-of-the-art methods that are based on different approaches. These include methods that use one (or a combination) of two-stream [34, 35, 89, 349, 131] or multi-stream [46] models, attention modules [34, 35, 347, 131], 3D CNN [36, 131], multi-task learning [348, 36], and methods that exploit hand posture data [351, 348, 36, 350].

The results in Table 4.1 demonstrate that SparNet achieves state-of-the-art performance in all benchmarks and experimental protocols. This indicates that the motion clues induced in the (single) appearance stream by the MP tasks were effective in improving the discriminative capabilities of the final embeddings, surpassing the performance of explicit optical flow information or 3D CNNs, and without the need for additional attention modules or supervised information. These results also confirm the findings from the ablation experiments, highlighting the effectiveness of the combined MP tasks and the relatively lower contribution of OFR to the overall performance of SparNet. As for the MP tasks considered, these results confirm the ablation ones, i.e., the optimality of the combined MP tasks and the lower contribution of OFR.

Experiments on EK In this work, due to the unavailability of test labels for the EK challenge at the time of submission, we followed the experimental protocol proposed in [36], which mirrors the "unseen" kitchen split of the EK challenge. We acknowledge that the EK challenge is extremely challenging, and many existing approaches involve complex architectures, combinations of 3D convolutions, prior supervised knowledge, multi-stream approaches, and ensemble methods, making it difficult to isolate the individual contribution of each component [352, 353, 36, 354, 47, 355].

In contrast, our approach focused on making minimal adjustments to the architecture to meet the specific requirements of the EK challenge. We added two separate fully connected (FC) layers at the end of the ConvLSTM for verb and noun prediction, and combined their outputs to define the action class. We used the average categorical cross-entropy loss for all tasks. Our rationale was to demonstrate the potential of a single, simple, yet effective approach in this challenging setting.

Table 4.5 compares our top-1 results using the combination of MS and OFC as the MP task (the most effective combination according to the results in Table 4.1) with the approach proposed in [36], which is the only other method available under the same experimental settings. As seen, the two approaches show comparable results, with SparNet achieving higher accuracy on verbs and lower accuracy on nouns and actions compared to [36]. These results partially contradict the findings from EGTEA+, where SparNet outperformed the other approach in all splits. There could be several explanations for these performance differences. First, EK contains unscripted video segments, which makes our uniform frame subsampling less optimal. Second, despite EK videos being longer than other datasets, we had to choose only 11 representative frames due to computational and memory constraints, which may not be optimal. Lastly, we note that [36] used a Multi-Fiber Network as their backbone,

EPIC-KITCHENS-55			
Method	Verb	Noun	Action
3DConv MTL [36]	49.31	**27.60**	**19.29**
SparNet-MS+OFC	**52.32**	26.01	16.95

Table 4.5 Top-1 accuracy (%) on the EPIC-Kitchens-55 validation set defined in [36]. The highest results are highlighted in **bold**.

which is a 3D CNN pretrained on Kinetics and may be better suited for video processing compared to our ResNet-34 backbone.

Interestingly, we observed differences in verb and noun accuracies between the two methods. The higher accuracy of SparNet on verbs could be attributed to the MP task helping our approach focus on capturing motion-related information, while the hand position regression task used in [36] helps their network focus on hand regions and implicitly on objects the hands interact with. The "naive" way of combining verb and noun predictions in our approach may explain the lower accuracy on action predictions. Moreover, our results suggest that the EPIC-Kitchens dataset presents more challenges compared to other available datasets, and our RGB-based approach struggles with generalizing to novel environments, such as unseen kitchens not included in the training data.

4.1.3 Conclusions

In this section, we introduced SparNet, a single-stream architecture for egocentric first-person activity recognition (FPAR). One of its key features is the joint learning of appearance and motion features through a set of self-supervised pretext tasks that estimate motion information from static input images. This results in a lightweight architecture that can be trained in a single stage and achieves state-of-the-art results on various publicly available datasets.

In conclusion, while acknowledging the limitations of optical flow information, we strongly emphasize the importance of incorporating multi-modal learning for the advancement of FPAR. In the upcoming sections, we will further investigate alternative modalities to substitute optical flow data, specifically focusing on improving the network's ability to accurately recognize hand-object interactions in the egocentric view. For instance, we will explore the role of audio signals in the egocentric vision context and investigate the potential of recent event data as a compelling compromise between wearable device efficiency and motion information encoding capability.

4.2 Multi-Modal Alignment – RNA

In this section, we focus on addressing the challenge of environmental bias or domain shift, which significantly impacts the ability of models to generalize to unseen scenarios. This limitation restricts the applicability of existing methods in real-world contexts. To tackle this issue, we present a novel approach for domain generalization in egocentric activity recognition. Our approach introduces a new multi-modal loss called Relative Norm Alignment (RNA) loss. The RNA loss re-balances the norms of the features extracted by the network across different domains, modalities, and classes. This ultimately improves overall accuracy on test data from an unseen "target" distribution. Furthermore, it can be easily extended to Unsupervised Domain Adaptation (UDA) setting by exploiting the availability of unlabeled target data during training. This is achieved by combining the RNA loss with a standard adversarial domain loss to further improve feature transferability and with an Information Maximization term to regularize predictions on target data. We present a comprehensive analysis and ablation of our method for both Domain Generalization (DG) and UDA settings and test our approach with different modalities. We also extend our analysis to other tasks, such as third-person action recognition, object recognition, and fatigue detection. The proposed approach achieves competitive or state-of-the-art performance on the proposed benchmarks, demonstrating the versatility of our method and its effectiveness in a wide range of applications.

As mentioned earlier, the use of wearable cameras and large-scale egocentric datasets has gained significant attention among researchers in the field of FPAR [34, 45–51, 10]. However, due to the camera's mobility with the observer, there is a higher degree of variability in factors such as illumination, viewpoint, and environment, compared to fixed third-person cameras. This variability negatively impacts the performance of egocentric action recognition models, and it is commonly referred to as "environmental bias" or domain shift [356]. It arises from a reliance of the model on the specific environment in which activities are recorded, hindering its ability to recognize actions when they are conducted in unfamiliar surroundings. To illustrate the impact of this problem, we present in Figure 4.3 the relative drop in model performance from the seen to an unseen test set of the top-3 methods of the 2019 and 2020 EPIC-Kitchens challenges. These results confirm that despite

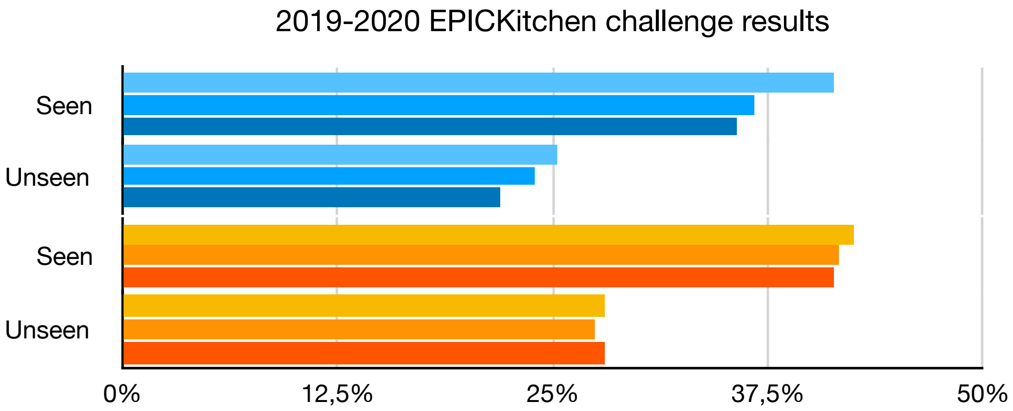


Fig. 4.3 Top-3 results of the 2019 [26] and 2020 [27] EK challenges, when testing on “Seen" and “Unseen" kitchens.

numerous efforts to find a specific solution for egocentric action recognition, the problem of environmental bias remains unsolved.

Recently, [37] addressed this issue by reducing the problem to an unsupervised domain adaptation (UDA) setting, where an unlabeled set of samples from the unseen test, called target, is available during training. While UDA can be effective in reducing domain shift, it may not always be practical in real-world computer vision applications. This is because it requires prior knowledge of the target domain, which may not be available beforehand, and because accessing target data at training time might be costly (or plainly impossible). Thus, we investigate an alternative solution that aims to enhance the network’s initial generalization ability while ensuring a seamless transition to the UDA scenario. This is known as domain generalization (DG) setting.

Inspired by the idea of exploiting the multi-modal nature of videos [37, 47], we make use of multi-sensory information to deal with the challenging nature of this particular setting. As mentioned in the previous section the RGB-optical flow are the two most widely utilized modalities in the egocentric action recognition[37, 88, 45, 46]. Furthermore, it has been shown that audio signals play an important role in egocentric action recognition, as they are naturally captured by most wearable devices [357]. Egocentric videos, in particular, contain rich sound information due to the close proximity of the sensors to the sound source during hand-object interactions, making audio a suitable modality for first-person action recognition [47, 92, 358]. For example, the action of "cutting" can exhibit visual and auditory

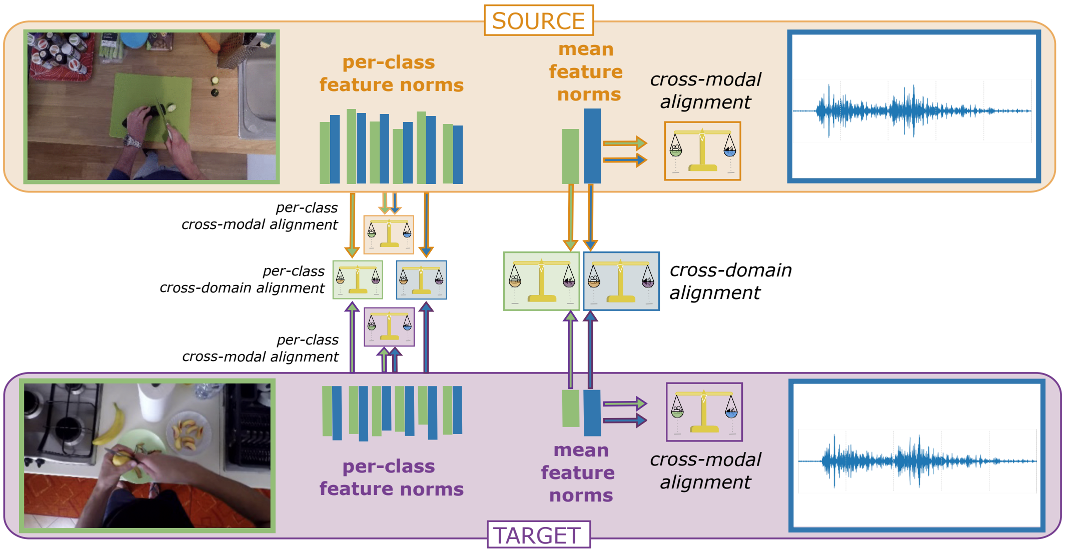


Fig. 4.4 Overview of Relative Norm Alignment (RNA) loss for RGB and audio modalities. Given visual and audio input from both source and target domains, we perform an alignment at feature level by re-balancing (i) the mean feature norms of visual and audio modalities (*cross-modal alignment*, $\mathcal{L}^{g}_{RNA}$), (ii) per-class mean feature norms of visual and audio modalities (*per-class alignment*, $\mathcal{L}^{c}_{RNA}$) and (iii) mean feature norms of source and target features independently for each modality (*cross-domain alignment*, $\mathcal{L}^{mod}_{RNA}$).

differences across domains, such as different types of cutting boards and food items being cut.

Despite multiple modalities could potentially provide additional information, the CNNs' capability to effectively extract useful knowledge from them is somehow restricted [359–363]. In our opinion, the origin of this difficulty is due to one modality being "privileged" over the other during training, as well as the main classifier tent to "privilege" features extracted from the source distribution. In particular, we observed that differences in the marginal distributions of different modalities do not only negatively affect the training process and lead to suboptimal performance, but also typically translate into discrepancies between the mean norms of their features. This imbalance in norms leads the network to "favor" the modality with the larger features, which prevents the model from fully exploiting the synergies and complementarities between modalities and reduces its generalization capabilities [364].

Motivated by these findings, we proposed to reduce such imbalance with a simple loss called *Relative Norm Alignment* (RNA) loss. In the DG setting, i.e., when the model does not have access to the target data at training time, this loss attempts to align the average norms of the different modalities to a common value. This objective also leads to successful transfer between source and target [365–367, 364]. In the UDA setting, i.e., when target data are available during training, RNA is defined as the sum of two domain-specific terms that aim to achieve a cross-modality norm balance on both source and target domains. While RNA encourages the alignment of all feature norms to a common value, it can also lead to imbalanced norms between classes which may penalize the ones with a smaller norm. Therefore, we include in the definition of RNA an additional component to enforce similar feature norms between classes intra- and inter-domain (Fig. 4.4, *per-class alignment*), which ultimately helps to improve overall accuracy.

In summary, the main contributions of this section are as follows:

- we bring to light the “unbalance” problem arising from training multi-modal networks, which causes the network to “privilege” one modality over the other during training, limiting its generalization ability;
- we introduce a new cross-domain and cross-modal loss, called the Relative Norm Alignment (RNA) loss. This loss function progressively aligns the relative feature norms and relative per-class feature norms of two or more modalities in DG setting and from source and target in UDA context;
- We conduct a comprehensive analysis and ablation study of our approach in both DG and UDA settings. We show state-of-the-art or competitive performance on all benchmarks and extend our analysis to multiple modalities and multiple tasks.

In the following, we detail the proposed Relative Norm Alignment (RNA) loss, which aims to mitigate the domain shift in multi-modal learning. We begin with a description of intuition and motivation, followed by a summary of the RNA description and formulation.

4.2.1 Intuition and Motivation

A common adopted strategy in the literature to solve the first-person action recognition task is to use a multi-modal approach [37, 47, 90, 92, 358, 88]. Despite the wealth of information of multi-modal networks w.r.t. the uni-modal ones, their performance gains are limited and not always guaranteed [359–363]. The authors of [359] attribute this limitation to overfitting and propose a solution by adjusting the loss value of each stream with different hyperparameters. However, this technique requires precise estimation, which is dependent on the task and dataset.

Norm *unbalance*. We hypothesize that during training there is an “unbalance" between the two modalities that prevents the network from learning “equally" from the two. Specifically, we believe that the network tends to “privilege" one modality over the other during training, while “penalising" the other. This hypothesis is also supported by the fact that the hyperparameters discovered in [359] differ significantly depending on the modality.

Several works highlighted the existence of a strong correlation between the mean feature norms and the amount of “valuable" information for classification [368–370]. In particular, the cross-entropy loss has been shown to promote well-separated features with a high norm value in [369]. Moreover, the work of [371] is based on the Smaller-Norm-Less-Informative assumption, which implies that a modality representation with a smaller norm is less informative during inference. All of the above results suggest that the $L2$-norm of the features gives an indication of their information content, and thus can be used as a metric to measure the unbalance between training modalities, classes and domains. After conducting an in-depth analysis of the feature norm during the training process, our hypothesis has been validated. The subsequent sections will provide detailed information on this validation. Leveraging this new approach to understanding the problem, we have proposed a novel multi-modal loss function called Relative Norm Alignment, which aims to improve the generalization capabilities of multi-modal architectures and address the domain shift issue.

4.2.2 Relative Norm Alignment Loss

The new loss function is suited to train a standard multi-stream architecture and it can be used in both DG and UDA scenarios. For simplicity, we will initially describe the RNA loss function in terms of two modalities, u and v, although the complete formula will be presented later. The input sample i consists of two modalities, $x_i = (x_i^u, x_i^v)$, which are fed into a features extractor (one for each modality m), as shown in Fig. 4.5. The features $f_i^m = F^m(x_i^m)$ are then processed by a classifier G^m,

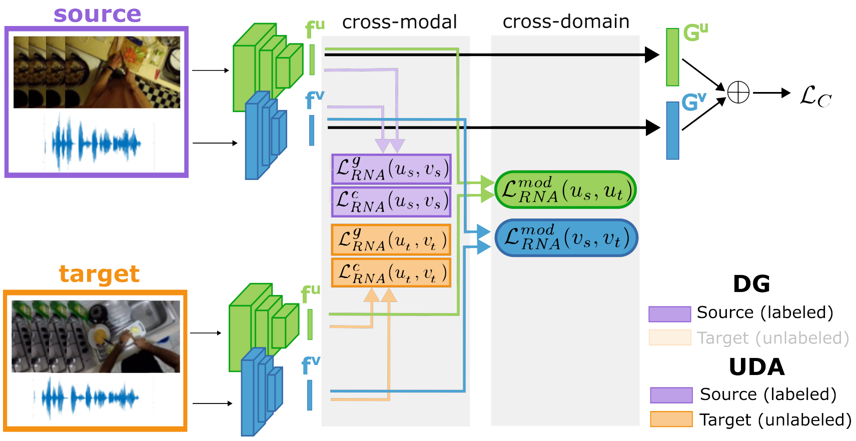

source
target
cross-modal
cross-domain
f^u
f^v
f^u
f^v
G^u
G^v
$\mathcal{L}_C$
$\mathcal{L}^g_{RNA}(u_s, v_s)$
$\mathcal{L}^c_{RNA}(u_s, v_s)$
$\mathcal{L}^g_{RNA}(u_t, v_t)$
$\mathcal{L}^c_{RNA}(u_t, v_t)$
$\mathcal{L}^{mod}_{RNA}(u_s, u_t)$
$\mathcal{L}^{mod}_{RNA}(v_s, v_t)$
DG
Source (labeled)
Target (unlabeled)
UDA
Source (labeled)
Target (unlabeled)

Fig. 4.5 Labeled source and unlabeled target samples from the modalities u (e.g., visual) and v (e.g., audio) are fed to the respective feature extractors. $\mathcal{L}_{RNA}$ aims to balance the relative feature norms of the two modalities, through a combination of the (domain-specific) cross-modal components ($\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$) and the cross-domain ones ($\mathcal{L}^{mod}_{RNA}$) in each u and v modality. In DG, only the components computed on the source are used.

which produces score predictions for the m-th modality of the i-th sample. Finally, a *late fusion* approach is used to combine the prediction scores from all modalities and obtain the final predictions. Our approach involves minimizing the following loss function:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{RNA} \tag{4.5}$$

Here, $\mathcal{L}_C$ represents the standard cross-entropy loss on the source data, while $\mathcal{L}_{RNA}$ is a novel loss term that is explicitly designed for DG settings and can be easily extended to UDA scenarios.

In particular, in DG setting the novel $\mathcal{L}_{RNA}$ is composed by the sum of two components:

$$\mathcal{L}_{RNA} = \mathcal{L}^g_{RNA}(u_s, v_s) + \mathcal{L}^c_{RNA}(u_s, v_s) \tag{4.6}$$

The $\mathcal{L}^g_{RNA}$ promotes a relative adjustment between the global norm of the modalities aimed at achieving an *optimal equilibrium* between them. This is expressed by the formula provided below, which defines the $\mathcal{L}^g_{RNA}$:

$$\mathcal{L}^g_{RNA}(u, v) = \lambda_g \left(\frac{\mathbb{E}[h(X^u)]}{\mathbb{E}[h(X^v)]} - \frac{\mathbb{E}[h(X^v)]}{\mathbb{E}[h(X^u)]} \right)^2 \tag{4.7}$$

where $h(x_i^m) = (\|\cdot\|_2 \circ F^m)(x_i^m)$ is the L_2-norm of m^{th} modality features of the i^{th} sample, $\mathbb{E}[h(X^m)] = 1/B \sum_{x_i^m \in \mathcal{X}^m} h(x_i^m)$ is the average norm for the m^{th} modality of the B samples composing the batch, and λ_g weights $\mathcal{L}_{RNA}^g$. To ensure that all features have the same dimension, we project them to a common shape using a fully connected layer when this condition is not met. The dividend/divisor structure of $\mathcal{L}_{RNA}^g$ (Eq. 4.7) promotes a relative adjustment between the global norm of the two modalities aimed at achieving an *optimal equilibrium* between the two. The square of the difference forces the network to take larger steps when the ratio of the two modality norms is too different, leading to faster convergence.

Although the $\mathcal{L}_{RNA}^g$ function is effective in enhancing the generalization ability of the multi-modal approach, the formulation presented in Eq. 4.7 does not account for the possibility that the *global* cross-modal alignment achieved by $\mathcal{L}_{RNA}^g$ may result in unbalanced norms between modalities at the class level. This can lead to a scenario where certain modalities are preferred over others when making decisions about specific classes. To address this issue, we introduce the intra-domain class constraint $\mathcal{L}_{RNA}^c$ as the second term in the $\mathcal{L}_{RNA}$ function. The $\mathcal{L}_{RNA}^c$ is designed to tackle the cross-modal norm imbalance at the class level, and is defined as follows:

$$\mathcal{L}_{RNA}^c(u,v) = \lambda_c \sum_{c=1}^{\mathcal{C}} \left(\frac{\mathbb{E}[h(X_c^u)]}{\mathbb{E}[h(X_c^v)]} - \frac{\mathbb{E}[h(X_c^v)]}{\mathbb{E}[h(X_c^u)]} \right)^2 \tag{4.8}$$

where λ_c weights the loss, and $\mathbb{E}[h(X_c^m)]$ denotes the average norm of the features of modality m for samples of class c, with C the total number of classes.

RNA for Domain Adaptation. In UDA setting the RNA is adapted as follows:

$$\begin{aligned} \mathcal{L}_{RNA} =& \mathcal{L}_{RNA}^g(u_s,v_s) + \mathcal{L}_{RNA}^g(u_t,v_t) + \\ & \mathcal{L}_{RNA}^c(u_s,v_s) + \mathcal{L}_{RNA}^c(u_t,v_t) + \\ & \mathcal{L}_{RNA}^{mod}(u_s,u_t) + \mathcal{L}_{RNA}^{mod}(v_s,v_t) \end{aligned} \tag{4.9}$$

Both the $\mathcal{L}_{RNA}^g$ and $\mathcal{L}_{RNA}^c$ are also minimized for the target data, in particular, to compute the $\mathcal{L}_{RNA}^c$ for the target, a pseudo-labeling strategy is used, assigning the target samples to classes. Furthermore, the last component of $\mathcal{L}_{RNA}$ addresses the problem of different norms in different domains by re-balancing the average and per-class norms of features in each modality across domains, so that the network can

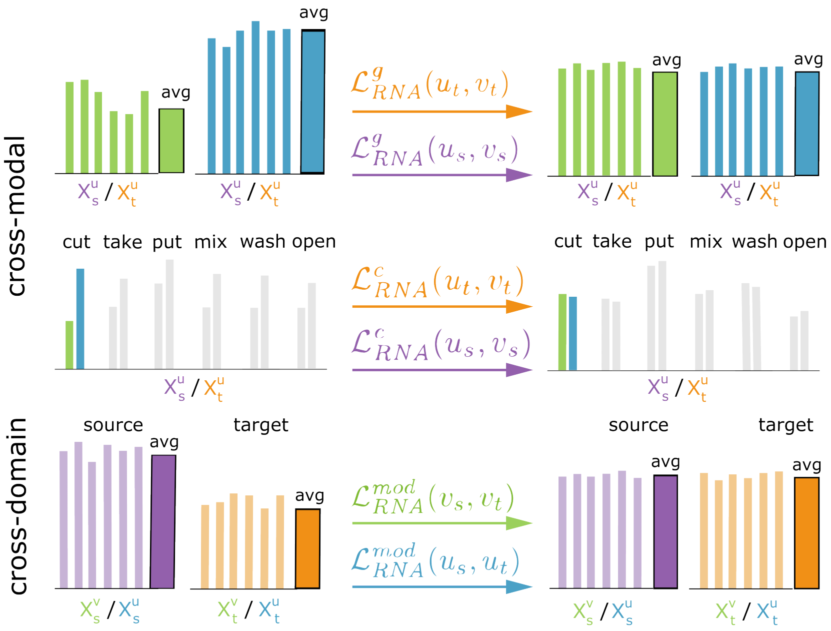


Fig. 4.6 *Individual effects* of the different components of $\mathcal{L}_{RNA}$ on the feature norms. Each diagram shows the norm per class of a single modality and a single domain (the color coding is the same as in Fig. 4.5). **First row:** $\mathcal{L}^g_{RNA}$ minimizes the overall average of the feature norms (larger bar on the right) of the different modalities (u and v for either source or target). **Second row:** $\mathcal{L}^c_{RNA}$ achieves the same goal at the individual class level (left: unbalanced norms, right: balanced norms thereafter). **Third row:** $\mathcal{L}^{mod}_{RNA}$ minimizes the difference between global and per class feature norms of the same modality across different domains. Here, diagrams represent the class and average norms for the same modalities (either u or v) before (left, unbalanced between domains) and after (right, balanced) the application of $\mathcal{L}^{mod}_{RNA}$.

focus on features that are more transferable between domains [367]. To this end, we include the following term in the formulation of RNA:

$$\mathcal{L}^{mod}_{RNA}(m_s, m_t) = \mathcal{L}^g_{RNA}(m_s, m_t) + \mathcal{L}^c_{RNA}(m_s, m_t) \tag{4.10}$$

where $m \in \{u, v\}$.

The individual contribution of the three losses is exemplified in Fig. 4.6. $\mathcal{L}^g_{RNA}$ globally aligns the norms of modalities for each domain. $\mathcal{L}^c_{RNA}$ aligns the norms of modalities per class for each domain. $\mathcal{L}^{mod}_{RNA}$ aligns the norms between domains, separately for each modality. Taken together, the three losses act synergistically. In DG, $\mathcal{L}^c_{RNA}$ supports the work of $\mathcal{L}^g_{RNA}$, which in turn facilitates the alignment of norms per class to a common value. The addition of $\mathcal{L}^{mod}_{RNA}$ in UDA helps the other two components to ensure that the average and per-class norms of the different modalities are also aligned between source and target.

Extension to multiple modalities The RNA objective in Eqs. 4.6 and 4.9 can be trivially extended to more than two modalities. In DG, the loss can be rewritten as:

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}(\mathcal{S}) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathcal{L}_{RNA}(i_s, j_s) \tag{4.11}$$

where i and j span the M modalities. Similarly, the UDA loss becomes:

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}(\mathcal{S}) + \mathcal{L}_{RNA}(\mathcal{T}) + \sum_{i=1}^{M} \mathcal{L}_{RNA}^{mod}(i_s, i_t) \tag{4.12}$$

where $\mathcal{L}_{RNA}(\mathcal{S})$ and $\mathcal{L}_{RNA}(\mathcal{T})$ are the loss in Eq. 4.11 for the source and target domains, respectively.

4.2.3 Learning objective in UDA

In addition to the loss defined in Eq. 4.9, to further improve the domain invariant properties of the features, we apply adversarial domain alignment [372, 373]. We follow the recipe used in other recent UDA work [189, 37, 374, 191], and introduce a classifier that predicts whether features are from the source or the target. This classifier is directly connected to the feature extractors via a Gradient Reversal Layer (GRL) [372]. The domain classification loss $\mathcal{L}_d$ is then multiplied by a weight λ_d and added to the total loss.

The loss we have introduced so far (i.e., the combination of $\mathcal{L}_{RNA}$ and $\mathcal{L}_d$) aims to improve the informative and domain invariant properties of the embeddings of the different modalities. However, these two loss components affect the feature extractors F^m and are not back-propagated through the classifier, which therefore only sees the source data and thus has no way to benefit from the target data. The result is that during training, the classifier focuses only on how best to integrate the multi-modal features to improve accuracy in the source domain, and completely ignores the *classification uncertainty* on target.

One approach commonly used in UDA to address this problem is to use a mutual information criterion [375] applied to the target data that not only minimizes the prediction uncertainty, but also promotes a uniform distribution of samples between classes. This is achieved through an Information Maximization (IM) loss defined as the difference between the average entropy of the outputs and the entropy of the average output:

$$\mathcal{L}_{IM} = -\mathbb{E}_{x\in\mathcal{X}_{\mathcal{T}}} \sum_{c=1}^{\mathcal{C}} p_c(x)\log p_c(x) + \sum_{c=1}^{\mathcal{C}} \bar{p}_c \log \bar{p}_c \tag{4.13}$$

where C is the total number of classes, p_c is the posterior probability for class c, and $\bar{p}_c$ is the mean output score for the current batch.

When we put all the pieces together, we train the model in the UDA setting to minimize the following loss:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{RNA} + \lambda_d \mathcal{L}_d + \lambda_{IM}\mathcal{L}_{IM} \tag{4.14}$$

where $\mathcal{L}_{RNA}$ is from Eq. 4.9 and λ_{IM} is the IM loss weight.

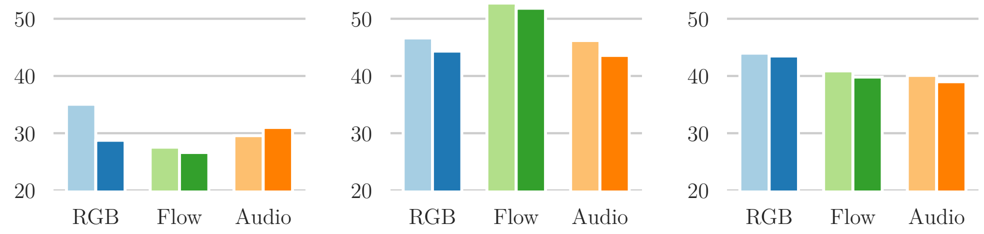
4.2.4 Experiments

In this section, we aim to verify the effectiveness of our proposed approach through an empirical evaluation on different multi-modal benchmarks corresponding to a variety of datasets and tasks. These range from action classification (on EPIC-Kitchens-55 [12], EPIC-Kitchens-100 [13], and UCF-HMDB [189]) to object recognition (on ROD [1]) and fatigue classification (on CogBeacon [376]). In particular, we have decided to include the CogBeacon dataset in our study to further validate the potential of our solution across various tasks and modalities beyond OR and AR. In the analysis, the results are obtained and presented as follows. When a dataset includes different domains, we optimized the models using the average accuracy over all the domain splits reported in the respective experimental protocol. Results were obtained using the same set of hyperparameters for all splits. Therefore, in the following, we excluded from evaluation the methods for which it was obvious (either from the description or from the available source code) that the hyperparameters were optimized for each split. For each benchmark, a description of the implementation details is provided, followed by an analysis of the results obtained.

4.2.5 Experiments on EK-100

EK-100 is the most extensive and diverse benchmark utilized in this work, for this reason, we choose to conduct an ablation study on it, enhancing the statistical significance of the findings. We follow the experimental setup proposed in [13], where the fine-grained nature of the dataset annotations combined with the large domain and temporal shifts between the source and target domains make the adaptation task very challenging. All the experiments in this section use all three modalities (RGB, Audio, and Flow) available in the dataset to test as well the model's ability to handle multiple modalities. The setting includes a validation split, for which labels are available, and a non-annotated test split. The results of this work are reported on the former, although previous work has also demonstrated the effectiveness of RNA on test data as well [377, 378]. Performance is evaluated in terms of Top-1 and Top-5 accuracy of verb and noun predictions and on the combination of the two predictions (action).

Implementation Details. RGB, Flow and Audio are processed following [47] by uniformly sampling 25 frames and 1.28 seconds audio segments along the action. During both training and inference, five of these samples are selected and fed to the network. Frame-level features $f_m \in \mathbb{R}^{25\times 1024}$ from each modality m are extracted using a TBN architecture [47] pre-trained on Kinetics [379] and fine-tuned on the source domain. During training, the feature extractors are frozen. Features are then fed to a multilayer perceptron and temporally aggregated using a TRN [91] module to obtain action-level features $f'_m \in \mathbb{R}^{1024}$. To account for the multi-task nature of this setting, we map the features into two components $f'_{m,v}$, $f'_{m,n} \in \mathbb{R}^{256}$, which we call *verb* and *noun features*. These are fed to two separate classifiers to obtain the modality logits for the verb ($y_{m,v}$) and the noun ($y_{m,n}$). Since this benchamrk includes a single source and a single target domain, the network is trained for action recognition by applying cross-entropy loss to the sum of *per-modality* logits. We extend RNA to work in this multi-task context by applying the alignment losses separately to the verb and noun features, immediately before the final classifier. Applying the RNA losses to these features ensures that the alignment effect provided by RNA is as close as possible to the classifier, which is heavily influenced by the feature norm values. The network is trained for 30 epochs using a batch size of 128



(a) Source only

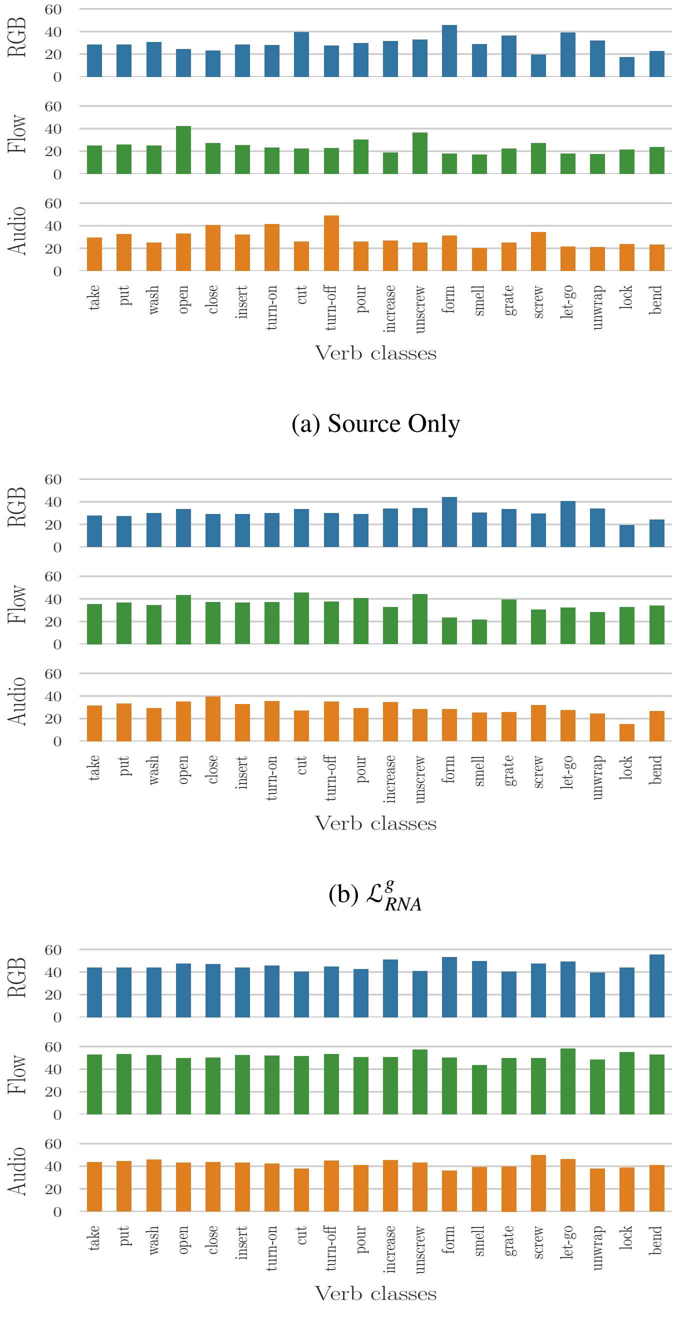
(b) $\mathcal{L}_{RNA}$ (DG)

(c) $\mathcal{L}_{RNA}$ (UDA)

Fig. 4.7 Verb feature norms across different modalities and settings (DG and UDA). Light (🟦 🟩 🟧) and dark colors (🟦 🟩 🟧) denote source and target validation domains, respectively. **(a)** In the *"Source Only"* setting, different modalities and domains result in unbalanced feature norms. **(b)** $\mathcal{L}_{RNA}$ in DG improves the alignment between different modalities, but leaves a gap between the source and target domains. **(c)** Finally, the contribution of $\mathcal{L}^{mod}$ in $\mathcal{L}_{RNA}$ reduces this gap in UDA, resulting in more consistent feature norms across different modalities and domains.

samples and SGD optimizer. The learning rate is initially set to 0.003 and decreased by a factor of 10 after epochs 10 and 20.

Effects of $\mathcal{L}_{RNA}$ on norm alignment. We begin by discussing the contribution of the components of the proposed $\mathcal{L}_{RNA}$ loss. Its goal is to mitigate domain shift issues by balancing the mean feature norms of the different modalities globally ($\mathcal{L}^{g}_{RNA}$), at the class level ($\mathcal{L}^{c}_{RNA}$), and across domains ($\mathcal{L}^{mod}_{RNA}$). In the following, we present the results of experiments in which these components are introduced incrementally.



(a) Source Only

(b) $\mathcal{L}^{g}_{RNA}$

(c) $\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA}$

Fig. 4.8 Feature norms of the top 10 most and least common classes from the target validation split of EPIC-Kitchen-100[13]. While $\mathcal{L}^{g}_{RNA}$ improves the alignment of different modalities, there is still an imbalance between classes. The addition of the per-class variant of RNA greatly improves this alignment, resulting in more uniform feature norms across different classes.

EPIC-KITCHENS-100			
Method	**CV**$_S$	**CV**$_T$	**CV**$_{S+T}$
Source Only	0.126	0.076	0.101
$\mathcal{L}^g_{RNA}$	0.089 (+29.3%)	0.121 (-59.8%)	0.103 (-2.1%)
$\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA}$ (DG)	0.075 (+40.4%)	0.098 (-28.7%)	0.081 (+20.1%)
$\mathcal{L}_{RNA}$ (UDA)	0.049 (+61.0%)	0.059 (+22.7%)	0.049 (+50.7%)

Table 4.6 Coefficient of variation for DG and UDA feature norms. CV$_S$, CV$_T$ and CV$_{S+T}$ are the CVs of the source, target and combined domain(s) respectively. For clarity, we also report the percentages of improvement with respect to the "Source Only" experiment.

Global alignment: a qualitative analysis. In Fig. 4.7 we report the mean feature norms for each modality. For simplicity, we will base our discussion on the verb feature norms, since the same observations apply to nouns. In particular, in Fig. 4.7 we show how the average norms of verb features for different modalities change on DG and UDA with the contribution of $\mathcal{L}_{RNA}$.

A preliminary qualitative analysis of the data presented in Fig. 4.7 shows that $\mathcal{L}_{RNA}$ in DG (Fig. 4.7b) leads to a better alignment of the average feature norms of the different modalities and to an overall increase of their values with respect to the "Source Only" (Fig. 4.7a). Recall that the norm formulation in Eq. 4.6 attempts to solve the alignment task at the batch level, and thus does not guarantee an exact alignment of all average norms. From Fig. 4.7b we can also observe the increase in the Flow norm, which is the lowest in "Source Only" (Fig. 4.7a). Flow has been shown to be the modality that is less sensitive to domain shift in egocentric action recognition [37], thus potentially guaranteeing greater generalization. This may justify the greater attention it receives by the network. We can also observe that the availability of the target data in UDA allows $\mathcal{L}_{RNA}$ to better align the norms of the three modalities across domains, which further improves the generalization capabilities of the final model, as shown by the improvements in accuracy reported in Table 4.7.

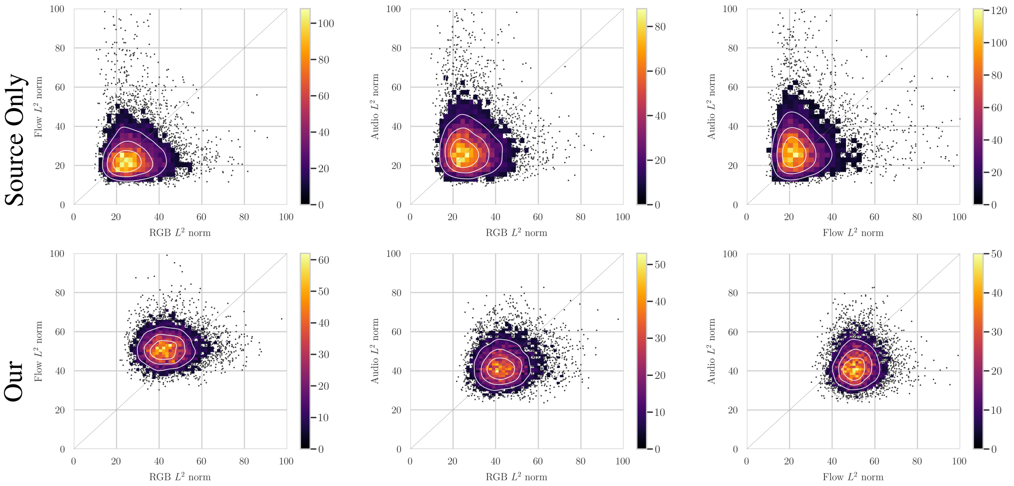
Global alignment: a quantitative analysis. To facilitate the assessment of the balancing effect of $\mathcal{L}_{RNA}$ between "Source Only", DG and UDA norms, we also introduce a quantitative metric. We use the *coefficient of variation* (CV) as a measure of the norm imbalance, with lower CVs indicating more balanced sets of values. CV is defined as follows:

$$CV = \frac{\sigma}{\mu}$$

where σ is the standard deviation and μ is the mean of the observed norm values. The CV values obtained are summarized in Table 4.6, where, for better clarity, we also report the percentage of improvement (%) with respect to the CV values of the "Source Only". As for the average feature norms in DG (Fig. 4.7b), we have a 40.4% decrease in CV compared to the "Source Only". It is interesting to note that the application of $\mathcal{L}^{g}_{RNA}$ alone only contributes to a 29.3% reduction of CV, highlighting the (positive) combined effect of $\mathcal{L}^{g}_{RNA}$ and $\mathcal{L}^{c}_{RNA}$. For the target domain in DG, we can observe that the imbalance between modalities increases (instead of decreasing) by 28.7%, which highlights the need for an alignment loss that works not only between modalities but also between domains. In UDA, the ability to use the target data contributes to a larger reduction in CV over the "Source Only" on both source (by 61.0%) and target domains (22.7%). When we consider the total imbalance (i.e., we calculate CV considering all source and target values together), CV shows an improvement of 20.1% in DG and of 50.7% in UDA. These values are reflected in progressively greater accuracy in the DG and UDA settings compared to the "Source Only" settings (Table 4.7).

Class alignment. For assessing the contribution of $\mathcal{L}^{c}_{RNA}$, we show in Fig. 4.8 the evolution of the verb norms of the ten most frequent and the least frequent classes in the DG settings. In the "Source Only" (Fig. 4.8a) the per-class mean features norms are largely unbalanced. While the exclusive use of $\mathcal{L}^{g}_{RNA}$ contributes to a better global balance of the modality norms, it has a small effect on the balancing of the norms per-class (Fig. 4.8b). On the contrary, when $\mathcal{L}^{c}_{RNA}$ is also minimized, we can observe a significant improvement of their alignment (Fig. 4.8c). These qualitative observations are also reflected in the CV metric computed on the class norms. Indeed, the use of $\mathcal{L}^{g}_{RNA}$ leads to a minor improvement in "Source Only" CV (37.8% and 19.5%, respectively, for source and target features) compared to that obtained by the combination of $\mathcal{L}^{g}_{RNA}$ and $\mathcal{L}^{c}_{RNA}$ (62.5% and 49.1%).

Overall effect on feature norms. To give further insight into the impact of $\mathcal{L}_{RNA}$, we show in Fig. 4.9 a scatter plot of the validation set in DG. This diagram is obtained by plotting the RGB, Flow and Audio feature norms of each sample in a 3D space whose axes are the norms of the three modalities. To make the plot easier to read, rather than using a single 3D representation, we present it as three separate



(a) Flow v. RGB (b) Audio v. RGB (c) Audio v. Flow

Fig. 4.9 Comparison of the feature norms before (top) and after (bottom) application of $\mathcal{L}^{g}_{RNA}$ and $\mathcal{L}^{c}_{RNA}$. The dots represent the samples in the validation dataset. The color bar on the right represents increasing density values. The original features, i.e. *"Source Only"*, show a wide range of values and an irregular shape, reflecting the misalignment between the features norms of the two modalities. The RNA loss re-balances the two, as evidenced by the more globular distribution while also shifting the average norms towards higher values.

sections along the coordinate planes defined by the feature pairs. The goal of these visualizations is to illustrate the changes in the shape of the resulting manifold.

It can be seen that the "Source Only" features are widely distributed and correspond to a manifold with a largely irregular shape. This is due to misalignment between the feature norms of the different modalities. When the $\mathcal{L}_{RNA}$ loss is applied, the manifold becomes more spherical and compact, reflecting the improvement in the alignment of the modality norms. It is also possible to note an increase in the average feature norm values that moves the manifold towards the upper right region of the 2D dimensional plots.

Effect of loss components. Table 4.7 details the contribution of the different loss components to the final performance in both DG and UDA settings. For better evaluation, we also show the average improvement in Top-1 accuracy of verb, noun, and action with respect to "Source Only" (Δ Acc). The combination of global and class components in DG ($\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA}$, Δ Acc. = 2.20) improves accuracy over $\mathcal{L}^{g}_{RNA}$ alone (1.36), showing that the combination of the two components effectively reduces domain shift. The ability to use target data in UDA boost the accuracy improvement

EPIC-KITCHENS-100				
Method	**Verb@1**	**Noun@1**	**Action@1**	**Δ Acc.**
Source only	46.79	26.79	18.29	-
DG				
$\mathcal{L}^{g}_{RNA}$	49.53	27.50	18.91	1.36
$\mathcal{L}^{g}_{RNA}+\mathcal{L}^{c}_{RNA}$	<u>50.75</u>	27.92	19.81	2.20
UDA				
$\mathcal{L}^{g}_{RNA}$	49.98	27.79	19.44	1.78
$\mathcal{L}^{g}_{RNA}+\mathcal{L}^{c}_{RNA}$	50.46	28.49	19.77	2.28
$\mathcal{L}^{g}_{RNA}+\mathcal{L}^{c}_{RNA}+\mathcal{L}^{mod}_{RNA}$	49.94	**29.48**	19.87	2.48
$\mathcal{L}_{RNA}+\mathcal{L}_{d}$	50.59	<u>29.38</u>	<u>20.04</u>	2.71
$\mathcal{L}_{RNA}+\mathcal{L}_{d}+\mathcal{L}_{IM}$	**50.82**	29.19	**20.05**	2.73

Table 4.7 Target Top-1 accuracy (%) achieved using various loss components. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>.

to 1.78 for $\mathcal{L}^{g}_{RNA}$ and 2.28 for $\mathcal{L}^{g}_{RNA}$ + $\mathcal{L}^{c}_{RNA}$), with $\mathcal{L}^{mod}_{RNA}$ further contributing to reaching an average improvement of 2.48.

As explained in Sec. 2.2, the learning objective in the UDA setting also benefits from two other losses, namely the adversarial domain loss $\mathcal{L}_{d}$, which aims to improve the transferability of features across domains, and the Information Maximization loss $\mathcal{L}_{IM}$, which aims to minimize the classification uncertainty between target classes. $\mathcal{L}_{d}$ provides a stronger improvement in this particular case (2.71), while $\mathcal{L}_{IM}$ has a minimal effect on the overall accuracy. However, we note that the mutual contribution of the latter two terms ($\mathcal{L}_{d}$ and $\mathcal{L}_{IM}$) also depends on the task and benchmark considered, as other experiments show more pronounced benefits for $\mathcal{L}_{IM}$.

Multi-modal adaptation capabilities. Another interesting question is whether the proposed method allows effective integration of multiple modalities in the final

EPIC-KITCHENS-100				
Method	**Verb@1**	**Noun@1**	**Action@1**	**Δ Acc.**
		RGB + Flow		
Source Only	44.80	25.35	16.33	-
Our (DG)	45.95	26.65	16.94	1.02
Our (UDA)	47.64	26.49	16.91	1.52
		RGB + Audio		
Source Only	39.91	24.18	14.84	-
Our (DG)	42.04	25.54	15.67	1.44
Our (UDA)	42.26	26.45	15.98	1.92
		Flow + Audio		
Source Only	45.11	21.98	15.37	-
Our (DG)	48.87	23.44	16.49	2.12
Our (UDA)	48.42	23.51	16.71	2.06
		RGB + Flow + Audio		
Source Only	46.79	26.79	18.29	-
Our (DG)	50.75	27.92	19.81	2.20
Our (UDA)	50.82	29.19	20.05	2.73

Table 4.8 Target Top-1 accuracy (%) on modality pairs on EPIC-Kitchens-100 [13]. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics.

EPIC-KITCHENS-100				
Method	**Verb@1**	**Noun@1**	**Action@1**	**Δ Acc.**
No Audio @ Test				
Source only	41.61	21.91	13.07	-
DG	44.03	24.44	14.89	2.26
UDA ($\mathcal{L}_{RNA}$)	44.08	24.77	15.25	2.50
No Flow @ Test				
Source only	30.58	20.33	10.63	-
DG	36.88	22.82	12.89	3.69
UDA ($\mathcal{L}_{RNA}$)	36.67	21.83	12.46	3.14
No RGB @ Test				
Source only	37.69	17.99	12.41	-
DG	46.70	18.92	13.53	3.69
UDA ($\mathcal{L}_{RNA}$)	46.51	19.37	13.55	3.78

Table 4.9 Target Top-1 accuracy (%) obtained through ablation experiments by dropping a modality at test time. All configurations are trained on all input modalities. At inference time, we simulate the loss of a modality which results in a large performance drop. RNA helps the model avoid focusing too much on individual modalities and is able to mitigate the performance drop. Δ Acc. is the average accuracy improvement for the verb, noun and action metrics.

decision and whether the use of multiple modalities also helps to improve the domain adaptation capabilities of the model.

Table 4.8 summarizes the results obtained comparing experiments with modality pairs and with all three modalities. It shows that the latter not only outperforms all other modality pairs in terms of results, but also shows better generalization properties, showing an improved delta compared to its "Source only" (2.73) compared to 2.06, the best two-modality improvement obtained with Flow + Audio. These results suggest that our method is effective in combining the different modalities to improve the overall accuracy and the generalizability of the features obtained.

Modality drop. In Table 4.9, we present an experiment to investigate the impact of modality imbalance during training. In particular, we examine the scenario in which a modality is "unexpectedly" lost at inference time without a training strategy accounting for this possibility. The basic idea of our approach is to help the model learn equally from the different modalities by integrating their contribution. While it is clear that the unexpected loss leads to a drop in accuracy, we can also expect that the effect of RNA is to make the model more robust to such a modality drop than the "Source Only" model, since the latter is less able to exploit the synergies between modalities and, thus, more vulnerable to dominant modalities. This expectation is confirmed by the results in Table 4.9, which show different but consistent effects on "Source Only" when different modalities are dropped at test time (i.e., large accuracy drops compared to the results in Table 4.7). At the same time, these results show that the balancing effect of RNA can potentially help the model reduce the impact of the lost modality, as it can take advantage of a better mutual contribution from the remaining ones.

Comparison with State-of-the-Art Methods. We continue our analysis presenting a comparison of the proposed method with the current state-of-the-art methods in both DG and UDA settings. We provide an overview of the comparisons and then discuss the obtained results.

Baselines. We compare our method with several multi-modal DG and UDA methods: MM-SADA [37], TA^{3}N [189], and CIA [212]. As for MM-SADA, the original approach works only with RGB and Flow modalities. Therefore, to integrate the Audio modality, we use two separate branches, one for RGB-Flow and the other for RGB-Audio modalities. The adversarial branch is applied individually to each modality. Finally, since our DG approach is primarily focused on improving the multi-modal learning capabilities of the model, we extend our analysis to include the Gradient Blending (GB) technique [359] as a DG comparison.

Results. In Table 4.10 results are given as Top-1 and Top-5 accuracy for verb, noun, and action. For each of the baselines, we also report the relative "Source Only" and its average improvement in terms of Top-1 accuracy. For the DG setting, comparisons are made using two approaches. The first is MM-SADA$_{SS}$, a modified version of MM-SADA that uses only the original self-supervised alignment task on the modalities applied to the source domain, and does not consider the adversarial alignment component of the method (which requires target data). The second approach is GB, which attempts to find an optimal blending of modalities according to their overfitting behaviour. Such a mixture is achieved by combining, with

EPIC-Kitchens-100

Methods	Verb@1	Noun@1	Action@1	Verb@5	Noun@5	Action@5	Δ Acc.
			DG				
Source Only	47.14	27.35	18.99	75.27	49.36	41.82	-
MM-SADA$_{SS}$ [37]	47.76	27.93	19.15	77.07	49.77	42.90	0.45
Source Only	50.27	29.04	19.96	81.74	52.14	46.74	-
GB [359]	50.18	**29.60**	20.26	**81.82**	52.57	**46.86**	0.26
Source Only	46.79	26.79	18.29	75.39	48.44	41.36	-
Our (DG)	**50.75**	27.92	19.81	80.64	51.37	45.33	2.20
Source Only†	49.81	28.55	19.77	81.10	51.90	46.22	-
Our† (DG)	50.20	29.31	**20.30**	81.58	**52.68**	46.76	0.56
			UDA				
Source Only	46.70	27.78	19.20	75.42	48.27	42.12	-
TA3N [189]	48.44	28.87	19.61	75.95	50.12	43.36	1.08
Source Only	47.14	27.35	18.99	75.27	49.36	41.82	-
MM-SADA [37]	48.44	28.26	19.25	77.56	50.59	43.41	0.82
Source Only	47.69	28.48	19.61	-	-	-	-
CIA [212]	48.34	**29.50**	**20.30**	-	-	-	0.79
Source Only	46.79	26.79	18.29	75.39	48.44	41.36	-
Our (UDA)	**50.82**	29.19	20.05	**80.89**	**52.18**	**46.04**	2.73

Table 4.10 Target Top-1 and Top-5 accuracy (%) on EPIC-Kitchens-100 [13]. Results are reported for the noun, verb and action metrics. Δ Acc. is the average Top1-accuracy improvement. †These experiments are trained using the cross entropy loss on both the fused logits as well as on the *per-modality* logits. The highest results are highlighted in **bold**, while other notable results are underlined.

appropriate weights, a cross-entropy loss for each modality and one for their fusion[1]. In terms of accuracy across different labels, results show that GB clearly performs best, while our approach is the runner-up and MM-SADA$_{SS}$ performs slightly worse. When we analyze the differences in "Source Only", we see that the one of GB is higher than ours, resulting in better delta accuracy values for our approach. This result seems to indicate that our method makes a greater relative contribution to reducing the domain shift. Nevertheless, the approach proposed in [359] is interesting and has some similarities with our approach in that it improves the balance between different modalities as a proxy for better classification accuracy. Therefore, we found it

[1] The original version of GB uses only RGB and Audio. The optimal weights for combining losses were taken from [27], and the weight for the missing component, i.e. Flow, was tuned appropriately for this work.

interesting to replicate our method with the "Source Only" of Gradient Blending, i.e., using multiple classification losses but without reweighting them. These additional experiments are marked with a † symbol. As can be seen in Table 4.8, the results are positive, allowing our method to achieve the best action accuracy and obtaining results competitive with GB (and also comparable with CIA, the state-of-the-art in UDA). However, we also emphasize that our standard solution solves the alignment problem with an adaptive approach that, unlike GB, does not depend on the model and the dataset used and requires only two hyperparameters, namely λ_g and λ_c.

In the experiments with UDA, we can see that, although being only the runner-up on actions, the delta accuracy improvement of our method is better than that of all other competitors, with results on the other metrics comparable to those of the other proposed baselines. At the same time, based on these results, we can observe that most of the improvements occur without accessing the target domain (and thus in DG), further highlighting the strong generalization advantage of RNA.

4.2.6 Experiments on EK-55

We adopt the experimental protocol of [37] and evaluate performance in a single-source setting ($D_i \rightarrow D_j$) on the three domains described in Sec. 2.4.1. Despite the small size of this setting compared to EK-100, it remains a highly valued and challenging benchmark in the field of egocentric action recognition due to the large domain shift between these domains and the unbalanced label distribution.

Implementation Details. As for the input, different sampling strategies are used to allow a fair comparison with the existing baselines. When using *dense sampling*, a clip of 16 consecutive frames is randomly sampled from the video. When using *uniform sampling*, 16 frames evenly distributed over the video are sampled. At test time, the same sampling strategy is used as in training, except that five clips are fed into the network instead of one, as suggested in [88]. In training, random clipping, scale shifts, and horizontal flipping are used for data augmentation, while in testing, only central cropping is applied. As for the aural information, we follow [47] and convert the audio track into a $256\times$ 256 matrix representing the log spectrogram of the signal. The audio clip is first extracted from the video and sampled at 24kHz. Then, the Short-Time Fourier Transform (STFT) is calculated with a window length of 10ms, a skip size of 5ms, and 256 frequency bands. For the Flow input, we use the same sampling strategy as for RGB. Both the RGB and Flow streams use an I3D model [94] as in [37]. Following [47], the audio feature extractor uses the BN-Inception model [380] pre-trained on ImageNet [330]. Each feature extractor produces a 1024-dimensional representation. Score logits for each modality are first computed using a single fully-connected layer and then fused by summing them. We train the network for $5k$ iterations using the SGD optimizer. The learning rate for RGB and Flow is set to $1e-3$ and reduced to $2e-4$ at step $3k$, while for Audio the learning rate is set to $1e-3$ and decremented by a factor of 10 at steps $\{1000, 2000, 3000\}$. The batch size is set to 128.

Comparison with State-of-the-Art Methods. In the experiments, we restrict our analysis to the RGB+Flow and RGB+Audio modality combinations, which are the ones recent work in the literature focus on.

Baselines. We compare our results with several state-of-the-art UDA methods. The first group (GRL [154], MMD [149], AdaBN [167], and MCD [381]) includes approaches originally developed as image-based methods and later adapted to work with video inputs. The second group includes more recent methods such as MM-SADA [37], the contrastive-based methods proposed in [56] and [55] (STCDA), and the recently published CIA [212]. In our comparison, we use the results reported in the original paper for each baseline.

Results. We begin by discussing the UDA results, which are summarized in Table 4.11. With respect to the RGB+Flow combination, and given the relevance of sampling strategies in the video context [382], we divided the results into different sections based on the sampling used for each modality (dense, D, or uniform, U). In particular, most baselines use D-D sampling and only CIA uses U-U sampling. In both cases, we compare the baselines to a version of our UDA method that uses the same sampling. It can be observed that CIA (uniform sampling), gives better results than the dense sampling-based methods. These results confirm the observation in [382] that uniform sampling usually allows the network to learn more information. We can also observe that our UDA approach yields state-of-the-art results for both samplings. To further validate the importance of sampling, we included experiments with a mixed sampling strategy (i.e., D for RGB and U for Flow) in table 4.11. Since there are no baselines with such sampling, we present only our results for “Source

EPIC-KITCHENS-55

Method	**Sampling**	**D1→D2**	**D1→D3**	**D2→D1**	**D2→D3**	**D3→D1**	**D3→D2**	**Mean**
RGB + Flow								
Source Only	D-D	42.00	41.20	42.50	46.50	44.30	56.30	45.47
GRL [154]	D-D	50.20	44.70	46.90	50.80	50.20	53.60	49.40
MMD [149]	D-D	46.60	39.20	43.10	48.50	48.30	55.20	46.82
AdaBN [167]	D-D	47.00	40.30	44.60	48.80	47.80	54.70	47.20
MCD [381]	D-D	46.50	43.50	42.10	51.00	47.90	52.70	47.28
DAAA [191]	D-D	50.00	43.50	46.50	51.50	51.00	53.70	49.37
MM-SADA [37]	D-D	49.50	44.10	48.20	52.70	50.90	56.10	50.25
Kim et al. [56]	D-D	50.30	46.30	49.50	52.00	51.50	56.30	50.98
STCDA [55]	D-D	52.00	45.50	49.00	52.50	52.60	55.60	51.20
Our (UDA)	D-D	50.84	47.14	48.86	54.38	50.6	58.43	**51.71**
Source Only	U-U	43.20	42.50	43.0	48.0	43.0	55.50	45.90
CIA [212]	U-U	52.50	47.80	49.80	53.20	52.20	57.60	52.18
Our (UDA)	U-U	52.84	47.49	54.41	54.11	55.53	61.64	**54.34**
Source Only	D-U	54.25	50.72	54.87	56.41	51.65	61.27	54.86
Our (DG)	D-U	56.00	50.39	56.25	56.37	56.73	61.63	56.23
Our (UDA)	D-U	57.33	52.84	57.19	56.78	57.27	62.03	**57.24**
RGB + Audio								
Source Only	D-D	39.03	39.17	35.27	47.52	40.255	49.98	41.87
GRL [154]	D-D	41.02	43.04	39.36	49.25	38.77	50.56	43.67
MMD [149]	D-D	42.40	43.84	40.87	48.13	41.46	50.03	44.46
AdaBN [167]	D-D	36.64	42.57	33.97	46.63	40.51	51.2	41.92
MM-SADA [37]	D-D	48.90	46.66	39.51	50.89	45.42	55.14	47.75
Our (DG)	D-D	42.55	41.77	42.73	51.09	42.63	54.24	46.21
Our (UDA)	D-D	46.65	47.22	46.18	52.30	44.04	56.18	**48.76**

Table 4.11 Target Top-1 accuracy (%) on EPIC-Kitchen-55 [12], using the evaluation protocol from [37], divided by modalities. Results are grouped by the sampling strategy used for a fair comparison. The highest results are highlighted in **bold**, while other notable results are underlined.

Only", DG and UDA. It can be seen that the "Source Only" method already achieves remarkable results (up to 3% better than our method with uniform sampling), which are further improved by our method in both DG and UDA (despite a much smaller difference with the "Source Only" than that obtained with other samplings). One possible explanation for the results obtained with this mixed sampling is that it implies a better exploitation of the properties of the two modalities. Indeed, dense sampling allows a better characterization of the (static) appearance information (RGB) over a short temporal range, while uniform sampling allows the use of a larger range to better capture the dynamic information conveyed by Flow.

Regarding the combination of RGB and Audio modalities, our UDA result is again the top performer (7% improvement with respect to "Source Only" and 1% improvement over the state-of-the-art method), confirming the potential of our method even when combining heterogeneous modalities.

Finally, we discuss the results from DG for both modality combinations. For RGB+Flow, we report the results with mixed sampling (D-U), i.e., teh sampling that gives the best results. In this setting, the results improve by up to 2% and 5% the "Source Only" of RGB+Flow and RGB+Audio, respectively, and the obtained performances are also close to those of the UDA setting (-1.01% and -2.55% for RGB+Flow and RGB+Audio, respectively). Although no other DG methods are available for comparison in this context, these results show that the DG setting can compete with several existing UDA methods that benefit from target data during training.

4.2.7 Experiments on UCF-HMDB

We conducted experiments on UCF-HMDB to evaluate the performance of our method on third-person action recognition. We follow the same experimental setting proposed in [91], which includes the U → H and H → U shifts in a multi-modal setting that includes the RGB and Flow modalities available with this dataset.

Implementation Details. For both RGB and Flow, the training input consists of 16 consecutive frames with resolution 224 x 224 pixels. In testing, we use five clips uniformly sampled across the video. The backbone for both RGB and Flow is an I3D pre-trained on Kinetics [379]. The learning rate is set to 0.01 and we train the

UCF-HMDB			
Method	**U→H**	**H→U**	**Mean**
Source Only	82.8	90.7	86.7
MM-SADA [37]	84.2	91.1	87.6
Source Only [55]	82.8	89.8	86.3
STCDA [55]	83.1	92.1	87.6
Source Only [56]	82.8	90.7	86.7
Kim et al. [56]	84.7	92.8	88.7
Source Only [212]	86.1	92.5	89.3
CIA [212]	88.3	94.1	91.2
Source Only (conc) [212]	85.8	93.5	89.5
CIA (conc) [212]	90.6	94.2	**92.4**
Source Only	83.6	94.1	88.9
Our (DG)	83.3	94.9	89.1
Our (UDA)	86.4	94.3	90.4

Table 4.12 Target Top-1 accuracy (%) on UCF-HMDB on RGB+Flow combination. The highest results are highlighted in **bold**, while other notable results are underlined.

model for 20 epochs with batch size of 32. We use SGD as the optimizer with a momentum of 0.9 and a weight decay of 10^{-7}.

Comparison with State-of-the-Art Methods. We compare our method to the other multi-modal UDA approaches discussed in the preceding paragraphs (MM-SADA [37], STCDA [55], the method of Kim et al. [56] and CIA [212]). To allow a fair comparison, all multi-modal results are based on the same backbones and the same pre-training.

Results. We present the classification accuracies of our method and several baselines in Table 4.12. Again, to ensure a fair comparison, we report the results of the "Source Only" model from the original paper for all baselines. In absolute terms, our approach under the UDA setting is the second best in terms of accuracy results, performing better than all baselines except CIA. However, we highlight the better "Source Only" result of CIA, which we found difficult to reproduce. Moreover, unlike our approach, the method proposed by CIA, i.e., improving spatial consensus between modalities, is not easily extensible to modalities other than RGB and Flow (as can be seen in [212] for the integration of Audio modality in EK-100). In terms of domain shift reduction, our approach achieves performance gains on "Source

SynROD ⟹ ROD		
	RGB + D	**RGB + E**
Source Only	47.70	49.19
GRL [154]	59.51	55.11
MMD [149]	62.57	62.39
SAFN [367]	62.40	<u>66.87</u>
Entropy [383]	63.12	66.23
Relative Rotation	<u>66.68</u>	66.68
Our (DG)	50.06	50.61
Our (UDA)	**82.36**	**78.52**

Table 4.13 Target Top-1 accuracy (%) on SynROD→ROD. The highest results are highlighted in **bold**, while other notable results are <u>underlined</u>.

Only" comparable to those of other methods. For example, the gains for MM-SADA, STCDA, Kim et al. [56], and CIA are 0.9%, 1.3%, 2%, and 1.9%, respectively, while our approach has a gain of 1.5%, with a maximum improvement by up to 3% on the U → H shift.

4.2.8 Experiments on ROD

We follow the experimental protocol already introduced in Chapter 3 (Section 3.2.1) for RGB-depth modalities, and the one in 3.3 for RGB-event. The studied shift is a synthetic-to-real domain shift, with synthetic source data and real target data (SynROD → ROD). RGB and depth modality in the synthetic domain are rendered, while events in the synthetic domain are simulated using ESIM [44].

Implementation Details. Event representations, depth images and RGB images are pre-processed and augmented during training following the procedure in [3]. Depth images are colorized with surface normal encoding, as in [18]. Input images are normalized with the same mean and variance used for the ImageNet pre-training, while we kept event representations un-normalized as this provided better performance. We use voxelgrid representation for events with 9 bins as in [305]. All backbones are implemented using ResNet-18 [339], and initialized with values obtained by pre-training the networks on ImageNet [330]. All the parameters of the network, including the pre-trained parameters, are updated during training, as in [3].

We train all network configurations using SGD as optimizer, batch size 64 and weight decay 0.003.

Comparison with State-of-the-Art Methods. The results for the comparison between SynROD and ROD using RGB, depth, and event modalities in both DG and UDA settings are presented in Table 4.13.

Baselines. We compare our results with standard image-based UDA methods, namely GRL [154], MMD [149], SAFN [367] and Entropy [383], which we extend to operate on multiple modalities. We also compare with our Relative Rotation approach.

Results. The comparison with the baseline UDA methods show that our method significantly outperforms all of them, with an improvement of up to 20% for the RGB+Depth combination and up to 10% for the RGB+event combination in the UDA setting. In the DG setting, the improvement with respect to the "Source Only" is quite small for both modality combinations. This can be attributed to the large gap between the distribution of source and target features in the synthetic-to-real setting, which leads to low generalization capabilities when the target is not available. On the contrary, better performance are obtained when there is a possibility to adapt to the target during training.

4.2.9 Experiments on CogBeacon

We follow the experimental protocol in the supplemental of [214], evaluating the performance in the single-source setting ($V_i \rightarrow V_j$) using three different domains (V1, V2, and V3), for a total of six splits.

Implementation details. The EEG signals are characterized using a total of 24 temporal and spectral features (see [376] for details). The face data are represented as a vector combining the average values of the face data and their standard deviation, yielding a total of 280 values. Both backbones are implemented by three 1D convolutional blocks with kernel size three and stride one, followed by a MaxPool layer and ReLU as the activation function. The output channels are 16, 32 and 64 for EEG signals and 8, 16 and 32 for the face keypoint model. The latter ends with a fully connected layer with an output of 64 to match the output of the EEG backbone.

CogBeacon			
Results from [214]		**Ours**	
Source Only	63.64	Source Only	61.80
MDANN [210]	66.83	SAFN [367]	64.01
MCD [381]	66.75	GRL [154]	64.24
CBST [384]	67.71	MM-SADA [37]	65.40
MM-SADA [37]	67.75	MMD [149]	65.58
DLMM [214]	**70.47**	Our (DG)	62.64
		Our (UDA)	**68.63**

Table 4.14 Target Top-1 accuracy (%) on CogBeacon. The highest results are highlighted in **bold**, while other notable results are underlined.

Predictions for each modality are computed with a single FC layer followed by a LogSoftmax. We train the model for 90 iterations using Adam as the optimizer. In all experiments, the learning rate was set to $1e-3$ and decremented by a factor of 10 at step 70.

Comparison with State-of-the-Art Methods. Table 4.14 presents the results of our proposed method compared to state-of-the-art methods using the Cobegon dataset, which consists of multi-modal data comprising EEG signal and face key-points.

Baselines. We compare our results with those in [214] (in particular with DLMM, the Differentiated Learning framework proposed in [214]) and with those obtained in our experiments with different UDA methods, namely SAFN [367], GRL [154], MMD [149], and MM-SADA [37]. These two lists of results are presented separately in Table 4.14 because the number of samples does not match that used in [214] (i.e., we have 2,240, 2,432, and 2,300 for domains V1, V2, and V3, respectively), a difference that we could not clarify with the authors.

Results. The results show that in UDA (68.63%) our method largely improves the "Source Only" baseline (61.80%). Although the improvement in DG (62.64%) is not as substantial, we think it still indicates a potential for softening domain shift. It is noteworthy that our method outperforms the other UDA methods used for comparison. Among them, MMD performs best with an accuracy of 65.58% (i.e., 3.05% less than our method). However, when compared to the results reported in [214], our method falls behind DLMM, which achieves an accuracy of 70.47%, significantly better than our results. Nevertheless, a comparison of the performance of DLMM and our method with the corresponding “Source Only” shows interesting results. DLMM achieves a 6.83% improvement over the “Source Only” baseline, and our method shows a 7.03% improvement, which can be considered equivalent. We would like to highlight that our approach is characterized by its simplicity compared to DLMM, which requires multiple training stages and uses a more complex curriculum learning approach with teacher/student models for different modalities. On the contrary, our method requires less computational resources and is easier to train, making it a more practical option for real-world applications.

4.2.10 Limitations

The proposed approach provides interesting performance in many cases, as shown by our experiments with a variety of tasks and scenarios. While the simplicity of the method is certainly a strength, it may be viewed as less effective when compared to methods developed and tuned for a specific task and benchmark. However, we believe that this limitation does not undermine the overall effectiveness of the proposed approach, as it provides a viable alternative for addressing various tasks without requiring significant computational resources or architectural changes.

Another limitation we observed arises from the fact that in many real-world cases the data distributions are strongly unbalanced, leading to lower precision for the tail classes [385]. The literature shows how this imbalance translates into unbalanced norms of classification weights per class [386, 387] as well as unbalanced norms of features per class [388, 389]. In developing our method, we expected that balancing the norms per class could have a positive effect also in rebalancing the weights of the classifier for the tail classes. However, our experimental results show that this effect is not present. This opens up possibilities for future developments to incorporate this objective into RNA as an additional component that rebalances the weights of the classifier.

4.2.11 Conclusion

In this work, we proposed a novel approach to address the problem of multi-modal domain adaptation. Our method starts from the observation that differences in the marginal distributions of modalities can negatively affect the training process, cause suboptimal performance, and induce imbalances in feature norms that limit the model's ability to exploit the synergies and complementarities between modalities. To tackle these issues, we proposed a Relative Norm Alignment (RNA) loss to balance the norms of the features extracted by the network in different domains and modalities and improve the overall accuracy. The method is applied in UDA with the combination of an adversarial domain loss and an information maximization term to improve feature transferability and regularize predictions in the target domain. Our experiments have shown that RNA can outperform or compete with several state-of-the-art approaches in a variety of multi-modal classification tasks, demonstrating its effectiveness and flexibility. Furthermore, the simplicity and lightweight nature of the proposed approach makes it easy to adapt to different architectures and contexts without requiring complex modifications. Future research will focus on further exploring the capabilities of RNA and integrating additional strategies to improve its performance in challenging scenarios.

4.3 From Pixels to Events – E^2(GO)MOTION

The objective of the last section of this chapter is to assess the applicability of event cameras as an alternative modality for recognizing motion information in the context of egocentric action recognition, where the limitations of optical flow information are well-established. As discussed in Section 2.3, event cameras are bio-inspired sensors that asynchronously capture pixel-level intensity changes in the form of "events". This study demonstrates that event data is a valuable modality for egocentric action recognition. To this end, we introduce N-EPIC-Kitchens, which is the first event-based camera extension of the large-scale EPIC-Kitchens dataset. Two strategies are proposed in this context: (i) directly processing event-camera data with traditional video-processing architectures (E^2(GO)) and (ii) using event data to distill optical flow information (E^2(GO)MO). On our proposed benchmark, we show that event data provides comparable performance to RGB and optical flow, without requiring additional flow computation at deploy time, and an improved performance of up to 4% *with respect to RGB-only information.*

The availability of wearable devices equipped with traditional frame-based cameras has led to a growing interest in egocentric vision, which is increasingly associated with visual RGB-data. However, recognizing actions in novel or unfamiliar environments remains a significant challenge for traditional RGB-based models [37, 54–57] that prioritize object textures and background cues, relying mainly on appearance-based information. To overcome these limitations, appearance-free modalities such as motion have been widely adopted in current egocentric vision systems. Unfortunately, computing motion through optical flow extraction from RGB frames is computationally expensive and impractical for real-time applications. Consequently, achieving state-of-the-art performance in real-world settings and finding a valid alternative to optical flow capable of unlocking cutting-edge methods for online predictions are still lacking in the field.

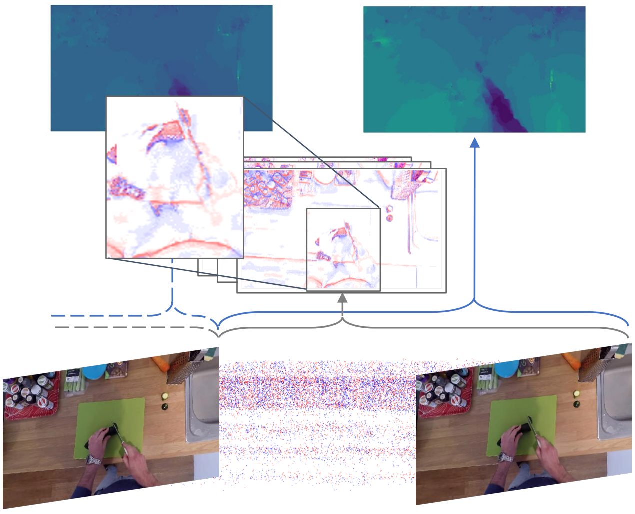


Fig. 4.10 **N-EPIC-Kitchens**: the first event-based dataset for egocentric action recognition. From RGB images, we generate a stream of events (bottom). Positive polarity is represented by red events, whereas blue events represent negative polarity. Events focus on motion, similarly to optical flow (top). With their low latency, high temporal resolution, and low-power consumption, event data are a perfect fit for egocentric action recognition.

Event-based cameras, on the other hand, have been shown to be particularly suitable for online settings [390]. Their high pixel bandwidth results in reduced motion blur, and the extremely low latency and low power consumption make these novel sensors particularly good in egocentric scenarios, where fast motion often impacts RGB-based systems negatively. Moreover, as they only convey differential information, event sequences reveal more information about the dynamic of the scene than its appearance, making them a valid alternative to RGB frames when learning to focus on motion. Still, despite these advantages, no prior research has looked at how to exploit their sensitivity to motion in egocentric vision, where these devices remain unused.

As a first step towards investigating the use of event data in egocentric action recognition, we propose a novel dataset called N-EPIC-Kitchens (Fig.4.10). It consists in the extension of the large-scale EPIC-Kitchens dataset [12] under the setup proposed in [37]. The latter is particularly appealing for both the availability of multiple environments (kitchens) and multiple modalities, i.e., RGB, optical flow, and audio. These features make it possible to analyze environmental bias and compare event data to well-established modalities. To further explore the potential of event data in egocentric action recognition, we introduce two approaches on N-EPIC-Kitchens that emphasize the intrinsic motion characteristics of event data. The first, which we call E^2(GO), consists in extending traditional 2D and 3D action recognition architectures with layer variations aimed at exploiting the motion-rich features of event data. The second, E^2(GO)MO, extends motion reasoning by distilling motion information from optical flow to event data. This is accomplished following a teacher-student approach that allows taking full advantage of expensive offline TV-L1 flow during training only, while avoiding its computation at test time. We summarize our contributions as follows:

- We present N-EPIC-Kitchens, the first dataset for event-based egocentric action recognition, which unlocks the possibility to explore event data in this context;
- We have evaluated N-EPIC-Kitchens using popular action recognition architectures, showing the performance of both single and multi-modal combinations with RGB and optical flow modalities. Moreover, we demonstrate the robustness of event data to environment changes;
- We have proposed two event-based approaches, E^2(GO) and E^2(GO)MO, which have been designed to highlight motion information captured by event data in egocentric action recognition.
- We show that event data can outperform RGB in challenging unseen environments and are competitive with them in known environments, suggesting that using event data is a viable option and more research should be performed in this direction.

4.3.1 N-EPIC-Kitchens

Event-based cameras have shown to be particularly efficient in egocentric scenarios as they focus on capturing only variations in the scene. This drastically reduces the amount of data to be processed and acquired, while also avoiding motion blur artifacts and providing fine-grained temporal information. Despite these advantages, the availability of freely accessible event-based datasets for human activity recognition is limited [29, 391, 282, 392]. Despite the field is actively working towards increasing their availability, as testified by the recent release of event-based versions of ImageNet [292, 393], relatively few datasets for human activity recognition are currently available. As shown in Figure 4.11, most of the available datasets focus

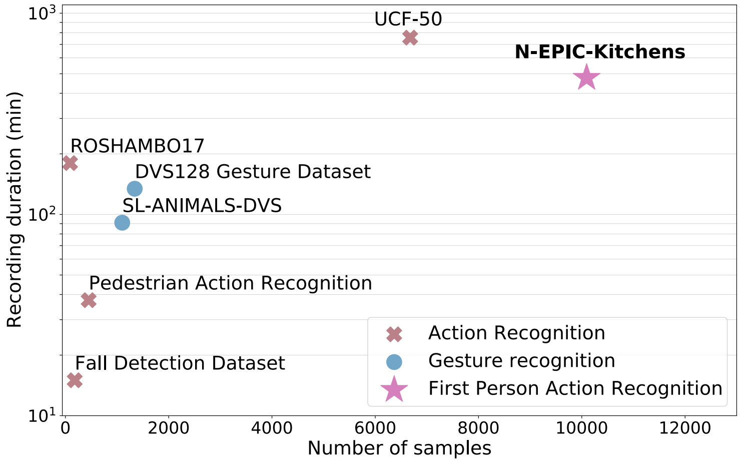


Fig. 4.11 N-EPIC-Kitchens *vs* existing event-based action classification datasets in the literature [28–32].

on action or gesture recognition in controlled settings where both the camera and background are static. None of the available datasets consider egocentric action recognition, preventing the use of event-based cameras in this scenario.

To showcase the efficacy of event-based cameras in online egocentric scenarios, as well as their ability to complement and compete with other modalities, we have expanded the EPIC-Kitchens (EK) dataset [12]. EK is a vast collection of egocentric action videos spanning multiple modalities and diverse environments. In line with the framework presented in [37], we selected the three largest kitchens from EPIC-Kitchens in number of training action instances, which we refer to as D1, D2 and D3, analyzing the performance for the 8 largest action classes, i.e., 'put', 'take', 'open', 'close', 'wash', 'cut', 'mix' and 'pour'. In the following, we first introduce the operating principles of DVS cameras. Then, we outline the approach used to generate N-EPIC-Kitchens and emphasize its benefits.

4.3.2 Event-Based Vision Data

Pixels of DVS cameras are independent and respond to changes in the continuous log brightness signal $L(\mathbf{u},t)$, differently from a standard RGB camera. An event is a tuple $e_k = (x_k, y_k, t_k, p_k)$ specifying the time t_k, the location (x_k, y_k) and the polarity $p_k \in \{-1, 1\}$ of the bright change (brightness decrease or decrease). An event is triggered when the magnitude of the log brightness at pixel $u = (x_k, y_k)^T$ and time t_k has changed by more than a threshold C since the last event at the same pixel, as described in the following equation:

$$\Delta L(\mathbf{u}, t_k) = L(\mathbf{u}, t_k) - L(\mathbf{u}, t_k - \Delta t_k) \geqslant p_k C. \tag{4.15}$$

Therefore, the output of an event camera is a continuous stream of events described as a sequence $\mathcal{E} = \{(x_k, y_k, t_k, p_k) | t_k \in \tau\}$, being τ the time interval.

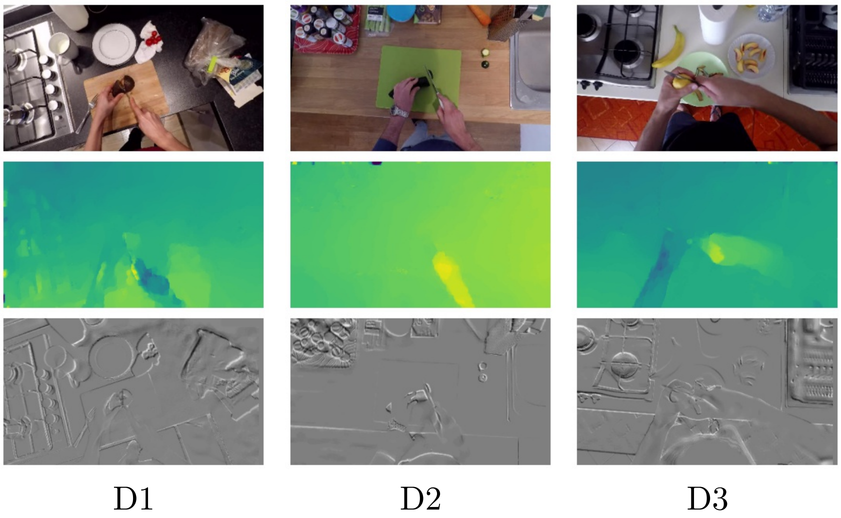


Fig. 4.12 RGB (top), optical flow (middle) and Voxel Grid representation (bottom) from the same action (“cut") on the three different kitchens (D1, D2, D3).

N-EPIC-Kitchens generation. We leverage a recent event camera simulator (ESIM) [44] to enhance the EPIC-Kitchen dataset with the event modality. ESIM allows us to produce events stream for a given brightness signal. Unfortunately, the original egocentric videos’ temporal resolution (milliseconds) is incompatible with one of the event cameras, which operates on a microsecond timescale. To address the RGB stream’s low temporal resolution, as in [294, 394], we use SuperSlow-MO [395]. It is a high-quality variable-length technique for frame interpolation that allows to increase the frame rate of the original video. Unlike previous frame interpolation approaches [396–399], SuperSlow-MO has the unique ability to generate frames with any temporal precision, which is ideal for our process. Finally, we use Voxel Grid [8], a frame-like event encoding technique, to convert sparse and asynchronous events to a tensor representation and enable learning with typical convolutional neural network architectures.

Challenges of Evaluating Event Data Assessing event data for first-person action recognition poses a fundamental challenge as the use of event modality in egocentric vision is entirely new. To overcome this challenge, we propose evaluating four different aspects of event-based modeling to establish a benchmark for this setting. We start by considering the importance of performance on both seen and unseen test sets, where *seen* indicates performance on the same kitchen on which training is performed, and *unseen* the performance obtained on a different one. This enables us to evaluate the upper bound performance of the modality and its ability to encode domain invariant features necessary for real-world applications. Then, as the performance of different modalities may greatly vary depending on the architecture used for processing [400], we benchmark events using three of the most accredited architectures in FPAR, namely TSM [90], TSN [121] and I3D [94]. To integrate event streams with off-the-shelf CNNs, we utilize a well-established procedure for converting event data into a frame-like representation that has been demonstrated to be efficient [295, 62]. Finally, we employ attention at the channel level to encourage the modeling of motion features.

Event Representation. Since event cameras produce sparse encodings of the scene, they must be converted into intermediate representations before processing. Several representations have been proposed, ranging from bio-inspired [244, 401, 402] to more practical ones. Frame-like representations are by far the most widespread methods as they can be directly used together with off-the-shelf networks. Among available ones [263, 264, 8, 402, 403, 269, 281, 271] we chose Voxel Grid [8] as it proved to be superior in cross-domain settings [295, 62]. This representation computes a B-channel image by discretizing time in B separate intervals:

$$\mathbf{x}^E(x,y,b) = \sum_{k=1}^{N} p_k k_b(b - t_k^*), \tag{4.16}$$

where b are the channels, t_k^* are the timestamps scaled into $[0, B-1]$, p_k is the polarity and $k_b(a) = max(0, 1 - |a|)$.

Backbone Architectures. To evaluate the performance of event data on different network designs, we analyze two popular 2D-CNN approaches, TSM [90] and TSN [121], and one 3D-CNN, I3D [94]. While TSN [121] only utilizes late fusion for temporal modeling, TSM [90] uses *shift modules* to exchange channel information across adjacent frames. On the other hand, I3D [94] is a pure 3D-CNN model that inflates filters and pooling kernels into the temporal dimension. In the literature, there is currently no clear preference for any particular technique, as certain modalities may react differently to different techniques. Therefore, we evaluate the performance of event data on all three techniques to assess which approach works best.

The Importance of Motion. Environmental biases are typically managed in egocentric vision systems by employing complementary, often appearance-free, modalities. Optical flow is generally the one performing the best in action recognition tasks [13, 12, 121], as (i) it helps focusing on the moving content, i.e., the action being performed, while (ii) preserving the edges of moving objects and (iii) ignoring background information. We argue in this section that, while event cameras are sensitive to moving edges and may ignore static information, they only record a portion of the three fundamental aspects of optical flow stated above. Event cameras can still detect events in the background due to camera movement, reducing their efficiency in filtering out less discriminative data. To solve this issue, we suggest using flow data to increase our capacity to filter out irrelevant information and improve action action recognition.

4.3.3 Learning from Motion

While a traditional RGB frame only encodes static information, event data frame-based representations also carry motion information on the channel dimension. Each temporal channel, in fact, captures the motion that happens in the blind-time between two normal frames of a video clip. We present two ways for enabling ordinary CNNs to leverage this information. The first, E^2(GO), explicitly models temporal relationships by providing channel operations that encourage motion reasoning. The second, on the other hand, employs a student-teacher technique known as E^2(GO)MO to encourage the network to extract motion properties during training by utilizing a pre-trained optical flow-based network. We detail the two approaches in the following.

E^2(GO): Event Motion In order to enable standard CNNs capture motion information from event data, we propose two simple but effective architectural variations, which improve the capability of extracting temporal inter-channel relations in 2D and 3D CNNs. We refer to them as E^2(GO)-2D and E^2(GO)-3D, respectively.

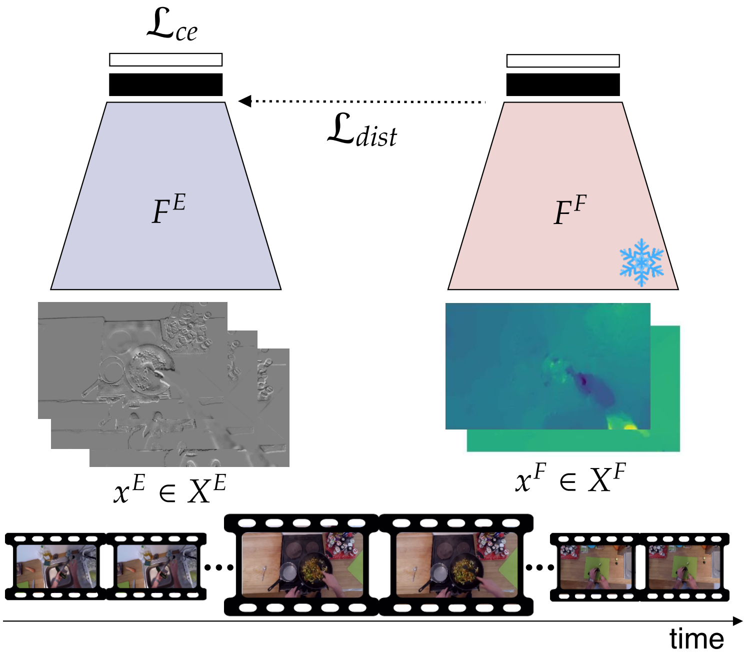


Fig. 4.13 Illustration of the proposed E^2(GO)MO. The input $\mathbf{x}^E$ and $\mathbf{x}^F$ from the event and flow modality are passed to the feature extractors F^E and F^F respectively. Information from the pre-trained teacher stream (frozen) F^F is distilled to the student stream F^E. The latter is trained with standard cross-entropy loss.

E^2(GO)-2D. A common practice in the literature is to extract temporal correlations at the video level by modeling dependencies between distinct frames [47, 90]. A feature of event representation is that the channel sequence captures continuous motion, describing micro-movements in the scene. This observation motivates us to extend the practice of modeling temporal relations to also learn short-range correlations between event channels.

Our approach involves utilizing *Squeeze And Excitation* modules [404] to enhance attention correlations between channels in 2D CNNs. The input to our model is an event volume $\mathbf{x}^E \in \mathbb{R}^{T\times H\times W\times F}$, where T represents the temporal dimension, $H \times W$ denotes the feature map resolution, and F corresponds to the number of channels. We refer to the features extracted from the i-th layer of the network as $\mathbf{f}_i^E \in \mathbb{R}^{T\times H_i\times W_i\times C_i}$. As a first step, we “squeeze" the spatial information content of $\mathbf{f}_i^E$ into a channel descriptor by performing feature aggregation along the spatial dimensions. It follows an “excitation" operator, which takes in input $\mathbf{z}_{sq}^E$ to produce an activation vector $\mathbf{s}$ to be used to scale $\mathbf{x}^E$. The scaling vector $\mathbf{s}$ is obtained from $\mathbf{z}_{sq}^E$ through two fully-connected layers with a bottleneck that down sizes C to C/r. Finally, $\mathbf{s}$ is used to re-weight $\mathbf{x}^E$, resulting in a new feature vector $\tilde{\mathbf{x}}^E$ to enhance discriminative motion features and discard the less informative ones. As a result, $\tilde{\mathbf{x}}^E$ encodes the relation dynamics between different temporal channels, effectively modeling the dependencies between them as a result of a self-attention function on channel dimension.

E^2(GO)-3D. Similarly, we propose to exploit 3D-CNNs' ability to process temporal information through a 3D kernel. Starting from the same input $\mathbf{x}^E \in \mathbb{R}^{T\times H\times W\times F}$, traditional 3D CNNs apply a 3D convolution on the (T,H,W,F) dimensions, resulting in an output of shape (T',H',W',C). We re-purpose the 3D convolution operator in this context to operate on $\mathbf{x}^E \in \mathbb{R}^{(F\cdot T)\times H\times W\times 1}$ by moving the channel dimensions on the temporal axis. This allows the convolution to capture the micro-movements present across the temporal channels of the event representation, which would have been ignored if processed only on the channel dimension.

E^2(GO)MO: Learning from Flow. Our objective is to train a network using both event and optical flow data, while eliminating the need to estimate the optical flow during testing. The input to our network is a multi-modal data $X = (X^E, X^F)$, where X^E represents the event modality, and X^F denotes the flow modality. We use F^E and F^F to refer to their respective feature extractors, and represent their resulting features with $\mathbf{f}^E = F^E(\mathbf{x}^E)$ and $\mathbf{f}^F = F^F(\mathbf{x}^F)$, respectively. Initially, we train the flow extractor F^F using cross-entropy loss between the true action labels $\hat{y}$ and the predicted labels y^F generated by a fully connected layer on top of F^F. Following this, we freeze the weights of the flow stream, and aim to transfer the knowledge from the pre-trained optical flow stream to the event stream. We first freeze the flow stream F^F and then train the event stream F^E by combining the standard cross-entropy loss with a distillation loss, which is defined as the L_2 distance between features $\mathbf{f}^E$ and $\mathbf{f}^F$:

$$\mathcal{L}_{dist} = \alpha||\mathbf{f}^E - \mathbf{f}^F||^2. \tag{4.17}$$

where α is a scaling hyperparameter. Such loss encourages features of the event stream to match those of the flow one, forcing F^E to mimic the behavior of F^F, and thus enabling the two to produce similar activations. Notice that we use optical flow data only during training and remove the teacher branch during inference, thus exploiting the advantages of this modality but effectively avoiding its computational complexity in prediction.

4.3.4 Experiments

In this section, we first introduce the experimental setup used, then we benchmark event data and validate the proposed E^2(GO) and E^2(GO)MO.

Experimental Setup

Input. Experiments with I3D [94] are conducted by sampling one random clip from the video during training and 5 equidistant clips spanning across all the video during test, as in [37]. The number of frames composing each clip is 16 for RGB and optical flow, and 10 for events. For TSN [121] and TSM [90] architectures, uniform sampling is employed by selecting five frames that are uniformly distributed along the video. During testing, five clips are selected per video, following the experimental protocol presented in [90]. The Voxel Grid representations are clipped between -0.5 and 0.5, and all data modalities are rescaled and normalized in accordance with the pretrained network associated with the architecture used. Standard data augmentation, as described in [88], is applied to all modalities.

Implementation and Training Details. Regarding the implementation and training details, the original implementation from [94] is utilized for I3D. On the other hand, TSN and TSM models are constructed using a BN-Inception [405] and a ResNet-50 [339] backbone, respectively. In the multi-modal experiments, a classic late fusion strategy is used, in which prediction scores from different modalities are summed and the error is backpropagated to all modalities. The models are implemented using PyTorch [406]. The optimizer used for training is SGD with momentum [407], with a starting learning rate of $\eta = 0.01$, a weight decay of 10^{-7}, and a momentum of $\mu = 0.9$. The networks are trained for a total of 5000 iterations with a learning rate decay to $1e{-}3$ at step 3000. All experiments are conducted using a batch size of 128 on 4 NVIDIA Tesla V100 16Gb GPUs. The hyperparameter $\alpha = 100$ is found to be optimal for the distillation loss.

Event Analysis. In Table 4.15 we show the performance of the event modality on the three selected action recognition architectures, varying the number of channels used for the voxel representation [8]. It can be observed that extracting 3-channels

EPIC-KITCHENS-55				
Model	**Voxel ch.**	**Testing**	**Seen acc**	**Unseen acc**
I3D	9	Clip	49.84	34.52
		Video	**52.50**	**36.24**
	3	Clip	53.75	35.90
		Video	**55.54**	**37.52**
	1	Clip	49.34	34.93
		Video	**51.29**	**35.05**
TSN	9	Clip	57.28	31.74
		Video	**58.98**	**32.52**
	3	Clip	58.81	34.65
		Video	**59.82**	**35.24**
	1	Clip	52.59	30.94
		Video	**54.54**	**31.87**
TSM	9	Clip	65.02	37.65
		Video	**66.39**	**38.71**
	3	Clip	64.38	37.75
		Video	**65.93**	**38.23**
	1	Clip	60.76	34.66
		Video	**62.46**	**36.45**

Table 4.15 Top-1 accuracy (%) achieved using I3D, TSN and TSM architectures depending on the number of channels for the event representation. The highest results are highlighted in **bold**.

Voxel Grid is the optimal choice and we used it in all the remaining experiments. In fact, it allows retaining the first ImageNet pre-trained convolution, which is otherwise trained from scratch when using a different number of channels. Indeed, the latter option is damaging on unseen domains. In fact, the first layers of the network are usually the ones that specialize the most on training data distribution [336], thus training them from scratch may lead the network to overfit on the training set, poorly generalizing on the unseen test. Instead, when exploiting pre-trained layers, the network can take advantage of robust low-level features.

In terms of performance on both seen and unseen test sets, the TSM model outperforms the I3D model, albeit only by a small margin. This can be attributed to the fact that the I3D model processes only a limited section of the video at any given time, thereby capturing only local features when trained at the clip level. Conversely, the TSM model works with full video frames, enabling it to capture global features.

EPIC-Kitchens-55

Input	Model	D_1	D_2	D_3	$D_1 \to D_2$	$D_1 \to D_3$	$D_2 \to D_1$	$D_2 \to D_3$	$D_3 \to D_1$	$D_3 \to D_2$	Seen	Unseen
RGB	I3D	53.67	61.12	60.70	34.50	35.70	34.94	36.46	33.93	38.37	**58.49**	35.65
Event	I3D	50.32	58.33	57.99	37.27	39.12	32.98	36.52	35.68	43.56	55.54	37.52
Event	E^2(GO)-3D	50.52	62.99	60.11	38.07	38.71	35.02	38.49	36.73	45.53	57.87	**38.76**
RGB	TSM	61.61	77.08	75.75	37.39	32.49	34.28	38.99	34.43	38.25	**71.48**	35.97
Event	TSM	56.86	72.43	68.49	28.73	34.00	37.09	42.30	42.27	45.02	65.93	38.23
Event	E^2(GO)-2D	56.58	70.03	69.60	34.98	35.16	38.21	47.80	41.71	44.13	65.40	**40.33**

Table 4.16 Top-1 accuracy (%) of event w.r.t. RGB on both I3D and TSM. Results are shown on all shifts, i.e., $D_i \to D_j$ indicates we trained on D_i and tested on D_j, and D_i means we trained and test on the same. E^2(GO)-3D and E^2(GO)-2D improvements are shown w.r.t. to their respective baselines, where no architectural variations are performed. The highest results are highlighted in **bold** on both seen and unseen for each backbone.

EPIC-Kitchens-55

Model	Streams	Pretrain	Seen (%)	Unseen (%)
I3D	Event	Kinetics-400 (R)	55.54	37.52
E^2(GO)-3D	Event	Kinetics-400 (R)	57.87	38.76
TSM	Event	ImageNet	**65.93**	38.23
E^2(GO)-2D	Event	ImageNet	65.40	**40.33**
I3D	Event+RGB	Kinetics-400 (R)	59.12	38.13
E^2(GO)-3D	Event+RGB	Kinetics-400 (R)	61.23	**41.85**
TSM	Event+RGB	ImageNet	71.88	39.92
E^2(GO)-2D	Event+RGB	ImageNet	**72.42**	40.61
I3D	Event+Flow	Kinetics-400 (R)	60.48	44.47
E^2(GO)-3D	Event+Flow	Kinetics-400 (R)	62.66	45.86
TSM	Event+Flow	ImageNet	72.26	46.89
E^2(GO)-2D	Event+Flow	ImageNet	**72.87**	**49.23**
I3D	RGB+Flow	Kinetics-400 (R)	62.07	44.56
TSM	RGB+Flow	ImageNet	**75.08**	**45.66**

Table 4.17 Top-1 accuracy (%) of the event modality when used in combination to stardard RGB and optical flow. The highest results are highlighted in **bold**.

As for the TSN model, its frame aggregation technique hinders the modeling of any temporal correlation, hence its inferior performance, which was expected. Thus, unless otherwise stated, we perform video-level analysis and evaluate the proposed approaches on TSM and I3D backbones in all of the following experiments.

Event *vs* RGB. In Table 4.16 we compare the behavior of the event modality against the RGB modality. Results show that the event modality outperforms RGB by a margin of up to 3% on unseen test sets. Indeed, it has been shown in the literature that appearance-based CNNs are biased toward texture, which causes them to underperform across-domain, but their robustness improves when shape-bias is

EPIC-KITCHENS-55													
Method	Model	**D1**	**D2**	**D3**	**D1→D2**	**D1→D3**	**D2→D1**	**D2→D3**	**D3→D1**	**D3→D2**	**Seen (%)**	**Unseen (%)**	**Mean (%)**
RGB	TSM	61.61	77.08	75.75	37.39	32.49	34.28	38.99	34.43	38.25	71.48	35.97	53.73
RGB + $\mathcal{L}_{dist}$	TSM	63.36	79.47	77.97	38.61	35.73	39.36	41.09	34.76	49.68	**73.60**	39.87	56.73 ▲+3
RGB + Flow	TSM	66.97	79.69	78.58	43.76	43.76	45.80	47.13	45.44	48.09	75.08	45.66	60.37
Event	TSM	56.86	72.43	68.49	28.73	34.00	37.09	42.30	42.27	45.02	65.93	38.23	52.08
Event	E^2(GO)-2D	56.58	70.03	69.60	34.98	35.16	38.21	47.80	41.71	44.13	65.40	40.33	52.87
Event	E^2(GO)MO-2D	61.38	75.83	75.08	39.77	37.19	44.71	51.03	47.01	53.73	70.76	**45.57**	**58.17** ▲+5.3
Event + Flow	E^2(GO)-2D	65.11	77.58	75.91	42.12	41.80	48.20	53.50	51.85	57.91	72.87	49.23	61.05

Table 4.18 Top-1 accuracy (%) of E^2(GO)MO w.r.t. the baseline on events (TSM) and E^2(GO)-2D. We compare E^2(GO)MO with the same approach on RGB to validate the choice of combining event and flow. The highest results are highlighted in **bold**, while underlined the best multi-modal.

increased [341]. Our hypothesis is that the event modality's superior performance is primarily attributed to its capacity to encode additional geometric and temporal information, which makes it more robust to variations in lighting and color, and thereby more invariant to domain shifts. This view is supported by the observation that RGB-based networks tend to overfit to domain-specific features, particularly in seen test sets. We remark that until now the event modality was still lagging behind RGB images in purely visual tasks, as reported by the recent release of N-ImageNet benchmark [292], where the best performing event architecture scores 48.94%, considerably below RGB's $>$ 90% accuracy [408–411]. Instead, our study demonstrates that the event modality exhibits a clear advantage over RGB in cross-domain scenarios, particularly within the context of egocentric vision. This highlights the importance of introducing shape and temporal biases to enhance the robustness of appearance-based convolutional neural networks and suggests that further research on the event modality has the potential to yield promising results in a variety of visual tasks.

E^2(GO). In Table 4.16, we present the results of our proposed models E^2(GO)-2D and E^2(GO)-3D. These models are designed to improve temporal correlations, enabling the network to emphasize informative motion features and suppress those that are not relevant to the action. Our experiments demonstrate that these models are particularly effective on unseen test sets. Specifically, E^2(GO)-3D achieved an improvement of up to 2% on the seen test set, while E^2(GO)-2D achieved results on par with the baseline TSM. These findings suggest that 2D CNNs, which rely heavily on visual signals and frame-based techniques, may not be as effective in changing environments, but they can still perform well in familiar environments. On the other hand, I3D, which uses convolutions across the temporal domain, is naturally more responsive to temporal correlations. By extending its temporal reasoning to micro-movements, our proposed models extract more discriminative features for the action, leading to higher accuracy even when testing on the same environment.

Multi-Modal Analysis. In Table 4.17 we illustrate the behavior of the event modality when used in combination of traditional ones, i.e., RGB and optical flow. When combined with RGB, it achieves an improvement of up to 7% on seen test sets and 3% on unseen ones. When combing event data with optical flow information, the best performance is achieved, improving event results by up to 7% on seen domains and 9% on unseen ones. This suggests that, while both event and flow encode motion, flow emphasizes the motion-relevant part, neglecting the scene or object affordances, while the event data maintain useful information about objects' shape (see Figure 4.12). For this reason, the event modality offers a potential advantage when combined with optical flow data than with RGB, which, instead, suffer on unseen domains due to its dependency on appearance. It is also worth noting that it outperforms standard RGB+Flow since event data is unaffected by negative features of appearance on unseen test set.

E^2(GO)MO. In Table 4.18 we illustrate the performance of E^2(GO)MO against an RGB-based TSM, which we proved to be the most robust architecture in the previous analysis. To prove our claim that the proposed distillation technique benefits from motion features, we also apply the same mechanism to an RGB-based stream, which we label in Table 4.18 with the RGB+$\mathcal{L}_{dist}$ entry. Both event and RGB benefit from the flow learning strategy, improving performance on unseen tests (+5.3% and +3% respectively), confirming the importance of motion information in real-world scenarios. However, E^2(GO)MO gains far more from the distillation loss $\mathcal{L}_{dist}$ than RGB, indicating that event data conveys more motion-rich features than RGB streams, thus proving our argument. Finally, we compare these two networks against their multi-modal upper bound performance, obtained exploiting the offline-computed optical flow also in prediction, namely RGB+Flow and E^2(GO)+Flow. Despite both are unable to reach their upper bound, E^2(GO)MO is much closer to E^2(GO)+Flow, and it even exceeds the multi-modal RGB-Flow performance. This result further motivates the use of event data over standard RGB in egocentric vision.

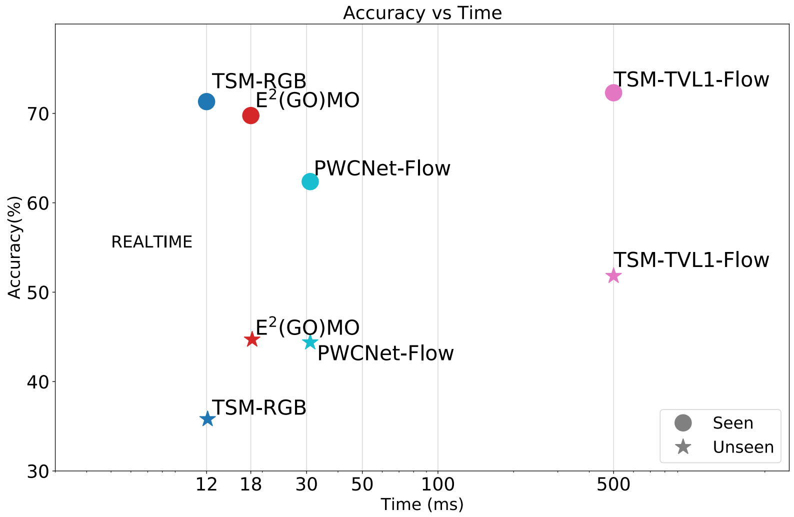


Fig. 4.14 Accuracy *vs* time of RGB modality, E^2(GO)MO, estimated PWCNet optical flow and TV-L1 optical Flow on seen and unseen scenarios for one clip evalutation.

Event *vs*. Optical Flow. Figure 4.14 depicts the trade-off between accuracy and average time per frame at test time, for both seen and unseen data. We evaluate the performance using two flow computation methods: TV-L1 flow, computed offline [59], and the one extracted from PWC-Net [412], which is one of the most competitive end-to-end CNN models for flow, providing an optimal balance between accuracy and time. The calculations were performed using an NVIDIA Titan RTX GPU, and we report both the input's computation and forward time, while ignoring data access time. In Figure 4.14, we also present the real-time capability of our proposed methods by highlighting the range of sufficient frame rates for a motion tracking system, using the threshold proposed in [413] as a reference. TV-L1 flow achieves higher accuracy than PWC-Net but at the cost of a longer extraction time (488 ms), which makes it unsuitable for online scenarios. On the other hand, the use of PWC-Net for online flow estimation leads to a significant drop in performance (up to 10% on seen tests and 8% on unseen tests).

Additionally, PWC-Net necessitates the execution of an additional network, increasing the parameter count ($\approx$ 40M) and requiring an additional fine-tuning stage. In contrast, we do not have to compute flow at test time, thus we can take full advantage of the more precise optical flow when distilling. Despite E^2(GO)MO does not explicitly use flow during inference, it still outperforms PWC-Net on seen tests (by up to 6%) and performs on par with it on unseen ones.

EPIC-KITCHENS-55

Stream	Model	Repr. Time (ms)	Seen (%)	Unseen (%)
RGB	I3D		**58.49**	35.65
Event	I3D	**6ms**	55.54	37.52
Event	E^2(GO)-3D	**6ms**	57.87	**38.76**
Flow (TV-L1)	I3D	488ms	58.47	43.40
RGB	TSM		**71.48**	35.97
Event	TSM	**6ms**	65.93	38.23
Event	E^2(GO)-2D	**6ms**	65.40	**40.33**
Flow (TV-L1)	TSM	488ms	73.23	53.98

Table 4.19 Top-1 accuracy (%) of RGB, Event and optical flow (TV-L1), along with their representation time, i.e., time to calculate the Voxel Grid for event, and extraction time for TV-L1 flow. The highest results are highlighted in **bold**.

Discussion and Limitations. Although the decision to simulate event data instead of creating a new first-person dataset was made to enable a direct comparison with established egocentric action recognition benchmarks [414, 12, 13], the inability to fully replicate event camera behaviors poses a significant challenge that may result in a domain shift between simulated and real-world event data, as discussed in Section 3.3.1.

Even though in the previous Section 3.3.1, our analysis primarily focused on the visual shift between simulated and real event data, considering it is an image-based classification task, it's crucial to recognize that another shift should be taken into consideration within the context of simulated event data. This additional shift arises from the contrast between the real high frame rate of the data captured with a DVS camera and the simulated one that is mainly derived from using SuperSlow-MO [395]. Isolating the impact of this particular domain shift (temporal domain shift) from the other one is a complex task. Moreover, it's important to note that in action classification tasks, the primary focus is on actions derived from human motion capabilities, which typically operate at a frame rate close to 20 fps. As a result, this temporal domain shift might not have a significant impact on performance. Nonetheless, it's reasonable to assume that domain adaptation or test-time adaptation techniques could help mitigate this gap. To gain insights into how this shift might impact performance, we can refer to the analysis performed in the task of Video Frame Interpolation (VFI) by [415].

The task of VFI consists of generating intermediate frames by inferring object motions in the image from consecutive key-frames, and in this specific work [415] the authors propose a method that uses the high-resolution capability of the DVS sensor in order to see in the blind time from two consecutive frames. In the analysis they provide results on the simulated and real parts of the High Quality Frames (HQF) dataset, showing a variation in performance from simulated and real data up to an average of 1.94 dB. This result confirms the existence of a domain gap between simulated and real event data even though it might not manifest at the same loss level as the one discussed in Section 3.3.1.

The intention to collect and analyze the impact of using real event streams, as demonstrated in [9], is an intriguing direction that can shed light on future research in the context of egocentric vision. This exploration is significant not only for understanding the challenges of egocentric vision but it could motivating advancements in future technologies. Indeed it's worth noting that, up to this point, there have been no smart glasses equipped with event cameras, making this research direction even more pioneering.

Furthermore, as shown in Table 4.19, TV-L1 optical flow still outperforms event data despite its high computational and time costs. This exceptional resilience to domain changes is primarily attributed to the fact that the algorithm for extracting TV-L1 optical flow partially filters out camera motion, resulting in cleaner motion data compared to unprocessed events. To address this limitation, another important direction for future work involves the use of motion compensation techniques commonly used with events [416], to remove redundant background noise.

4.3.5 Conclusion

In this section, we aim to investigate the potential of event data in the context of egocentric vision. Our motivation stems from the need to identify an alternative visual modality that can encode motion information more effectively than RGB data, while also having low computational requirements for online settings. Event-based cameras offer several advantages, such as high pixel bandwidth, extremely low latency, and low power consumption, which make them particularly well-suited for egocentric scenarios.

For this scope, we introduce N-EPIC-Kitchens, a novel event-based egocentric action recognition dataset. Leveraging the diverse modalities already available in the EPIC-Kitchens dataset, we conduct a comprehensive comparative analysis that highlights the significance of event data in the egocentric action recognition context. Based on these findings, we propose and evaluate two innovative event-based approaches, namely E^2(GO) and E^2(GO)MO, which prioritize motion information and achieve competitive results compared to the computationally expensive optical flow modality. Our extensive experiments demonstrate the resilience of event data and their effectiveness in action recognition settings, providing encouragement for the community to pursue further exploration in this direction.

Chapter 5

Conclusion

In this chapter, a thorough summary is presented that covers the significant outcomes and contributions highlighted in the thesis. Additionally, open issues are discussed, and potential areas for future research are outlined.

5.1 Summary

Despite the remarkable achievements of AI in recent years, there still exist disparities between the capabilities of intelligent systems and those of humans. Humans possess the remarkable ability to perceive and interact with the world in an efficient and adaptive manner. Our understanding of the world stems from a multi-sensory perception, enabling us to derive deeper knowledge beyond what can be extracted from a mere 2D representation. By integrating multiple modalities into AI systems, we can unlock new possibilities for enhancing accuracy, efficiency, and overall performance, and bridge the gap between human capabilities and machine performance in various real-world applications.

This thesis is dedicated to exploring the potential of multi-modal learning to enhance the capabilities of models in terms of accuracy, robustness, and adaptability. To achieve this goal, we delve into two specific research domains: Object Recognition discussed in Chapter 3, and First Person Action Recognition explored in Chapter 4. These research areas share the common goal of addressing the limitations inherent in relying solely on uni-modal approaches. Indeed, in Section 3.1, we introduce a novel end-to-end trainable model for OR using RGB-D data. This approach facilitates the interaction between RGB and depth channels, resulting in a final embedding that effectively captures the object characteristics. In the context of FPAR, accurately extracting motion information from videos poses a significant challenge for current models. To address this issue, Section 4.1 introduces SparNet, a network that simultaneously encodes appearance and motion information. This is achieved by incorporating a self-supervised pretext task that focused on motion segmentation. Notably, the multi-modal aspect plays a distinct role here. The optical flow, which captures motion information, is considered as non-real second modality since it is not directly captured by a sensor and its computation requires substantial time and resources. For this reason, our approach leverages optical flow solely during the training phase, while utilizing a single-stream network during the test phase that has significantly benefited from multi-modal learning.

Furthermore, the ability of models to generalize and adapt to new environments and tasks is a crucial aspect of advancing the capabilities of AI systems. In this regard, in Sections 3.2 and 4.2 we demonstrate how improving the interaction between modalities is essential for enhancing the models' generalization capabilities and subsequent adaptability, introducing novel multi-modal DG and UDA techniques.

In addition to exploring the benefits of multi-modal learning, this work also investigates the impact of new modalities. Specifically, Sections 3.3 and 4.3 delve into the opportunities and challenges associated with the utilization of event-based cameras. Specifically in Section 3.3, we have made significant progress in addressing the challenges associated with using simulated event data. This advancement has the potential to drive further research in the field of event vision, allowing for the exploration of new tasks even without access to real cameras. Furthermore, in Section 4.3, we introduce a pioneering event-based dataset specifically designed for fine-grained action recognition in a first-person perspective. This dataset facilitates the exploration of this novel modality for the aforementioned task and allows for direct comparisons with other established modalities. Finally, we present E^2(GO)MO, a strategy to adapt existing networks for action recognition to event-based data. This approach leverages the potential of encoding motion information offered by event-based data.

In conclusion, this thesis extensively explores multi-modal learning, focusing on two specific tasks. Significant advancements have been achieved in these contexts, in particular on the topics of UDA, and DG, and new possibilities are opened for the event data, particularly in the context of egocentric vision. Lastly, this thesis offers a crucial bridge that effectively connects different contexts characterized by diverse literature, application domains, and research communities. We firmly believe that establishing such connections not only facilitates the exchange of knowledge but also promotes future interdisciplinary collaboration, thereby contributing to the continuous advancement of these fields.

5.2 Open Issues

The progress made in this thesis has broader implications beyond the domains of OR and EAR. These advancements have the potential to benefit a wide range of tasks, including image-related tasks such as object detection and segmentation [417, 418], as well as other tasks within the field of egocentric vision [51, 419–421]. It is worth noting that the field of multi-modal learning continues to be an active area of research, as demonstrated by recent publications such as [422, 41, 423]. These recent contributions have made significant progress in pushing the boundaries of AI systems and highlighting the importance of ongoing research in the field.

Where does this thesis stand in relation to ongoing Large Language Models research? The recent years have witnessed a remarkable revolution in Large Language Models (LLMs), as they have demonstrated exceptional proficiency in natural language understanding, encompassing semantic comprehension, question answering, and text generation [424–426]. The application of LLMs to computer vision tasks has also seen significant interest. An example of this is CLIP [427], which involves the creation of image representations from scratch using a vast dataset comprising 400 million (image, text) pairs sourced from the internet. CLIP's noteworthy ability to generalize, particularly in zero-shot and few-shot scenarios, has been a significant breakthrough.

We firmly believe that the topics and methodologies explored in this thesis will serve as enduring references for future research in applying LLMs to computer vision tasks. Notably, many of the multi-modal methods introduced here possess architecture-agnostic characteristics. This means they are well-suited to work seamlessly with emerging models like CLIP. Furthermore, these methods have consistently yielded remarkable results at scale, as evidenced by their performance in prestigious competitions like EPIC-Kitchen.

Nevertheless, it is crucial to acknowledge certain limitations inherent in approaches like CLIP and similar models. Firstly, these models primarily focus on creating a text-based representation for uni-modal visual data, often sourced from the web. This narrow focus limits their consideration of other vital multi-modal and temporal information. In response to this limitation, the research community has made significant progress in adapting CLIP for video and various modalities [428–432]. This development underscores the potential validity of our research, especially in the domain of synthetic and simulated data, facilitating access to new annotated datasets and leveraging the properties of CLIP encoders for novel tasks and modalities, such as event data. Second, as regards the egocentric setup, despite reaching an impressive data scale, videos in those existing video-text pretraining datasets are often of 3rd-person views and may have been edited before posting on the Web. Yet, there is a noticeable domain gap between the existing video-text pretraining datasets and 1st-person view videos such as those videos directly captured by wearable cameras or smart glasses. Some progress to overcome this limitation and generate a Video-Language Pretraining that uses egocentric videos are made in some works like EgoClip [433] that exploits data taken from the large-scale dataset Ego4D.

In our recent work, which is not included in this thesis, we have demonstrated an important application of the CLIP model within the egocentric context. Specifically, our work showcases how CLIP can be effectively employed to extract agnostic affordance information from first-person videos. The concept of affordance is used already in neuroscience and cognitive psychology to describe a relationship between the actions and the environments in which they occur. This information holds relevance for various sub-tasks, including enhancing a model's capability to adapt to new environments. In conclusion, the substantial contributions to the realm of multi-modal learning introduced in this thesis are aligned with the evolving landscape of LLM research.

Multi-modal learning. In future research, there are several important directions to explore in the field of multi-modal learning. One key aspect is to explore scenarios where the availability and reliability of all modalities cannot be always guaranteed. This variability in the presence of input modalities poses unique challenges that require tailored solutions for effective handling. Developing robust methods that can adapt to variations in the number of input modalities will be crucial in real-world applications. Another interesting direction for future investigation is the development of methods that can dynamically activate and deactivate modalities or handle modalities with different sampling rates. This dynamic modality management capability is essential for optimizing the resource allocation in multi-modal learning systems. By selectively activating modalities based on their relevance or adjusting the sampling rates according to the task requirements, the overall efficiency and performance of the system can be improved. This becomes particularly important in resource-constrained environments where energy efficiency is a priority. Recent literature has already made significant progress in addressing some of these challenges, with notable contributions from works such as Alfasly et al. [434] and Woo et al. [435]. However, further research is necessary to refine and expand upon these approaches. Continued exploration of these directions will contribute to advancing the field of multi-modal learning and unlocking its full potential in various domains and applications.

Unsupervised Domain Adaption. One future research direction involves addressing the limitations of the DA setting. Currently, in DA setting we assume that the source and target domains share the same label space, known as Closed Set DA. However, in real-world scenarios, this assumption may not always hold. To overcome this limitation, exploring alternative settings such as Partial DA, Open Set

DA, and Universal DA would be valuable. These settings allow for variations in the label space between the source and target domains, including cases where there is partial overlap, full overlap, or no overlap at all.

At the same time, in the context of egocentric vision, the domain shift problem becomes more pronounced due to significant variations in data caused by rapid changes in the environment, perspective, and lighting conditions. Alternative approaches need to be explored to effectively address the domain shift issue. For example, methods such as Test Time Training or Adaptation (TTT or TTA) [436] or Continual Unsupervised Domain Adaptation (CUDA) can be investigated. TTT or TTA involves refining the model during the test phase, instead, CUDA focuses on adapting the model incrementally as new target domain samples become available over time, allowing the model to continuously learn and adapt to the evolving domain.

Event data. Our research has achieved remarkable results in utilizing event data for OR and FPAR. However, it is crucial to recognize that the current efforts, including those in the existing literature, largely involve adapting pre-existing architectures designed for standard image data to handle event data. To fully unlock the potential of event cameras, it is necessary to explore novel architectures explicitly tailored to this data. A promising avenue for further investigation is the development of architectures optimized specifically for this new modality, leveraging recent advancements in Neural Architectural Search (NAS) [437]. By employing NAS techniques, researchers can identify architectures that are well-suited for processing event-based data, thereby enhancing performance and computational efficiency.

Egocentric vision. In future research, an exciting direction is to explore ways to relax the constraints associated with egocentric action recognition and develop a real-time system capable of predicting ongoing actions as they occur and detecting their boundaries. This system should demonstrate adaptability to varying environments, effectively handling the dynamic nature of wearable sensor data. Additionally, it is crucial to consider the context of wearable devices and optimize the system to meet the requirements of small devices in terms of model size and computational efficiency. While some progress has been made in our previous work [438], which is not included in this thesis and in [439], further research is necessary to address the specific challenges related to action detection and the development of lightweight models. Advancements in these areas will contribute significantly to the development of a robust and efficient system for real-time action prediction in egocentric vision applications.

References

[1] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE international conference on robotics and automation*, pages 1817–1824. IEEE, 2011.

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[3] Mohammad Reza Loghmani, Luca Robbiano, Mirco Planamente, Kiru Park, Barbara Caputo, and Markus Vincze. Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition. *IEEE Robotics and Automation Letters*, 5(4):6631–6638, 2020.

[4] Blender. http://www.blender.org. Accessed: 2020-01-30.

[5] A. Aakerberg, K. Nasrollahi, and T. Heder. Improving a deep learning based rgb-d object recognition model by ensemble learning. In *IEEE International Conference on Image Processing Theory, Tools and Applications*, pages 1–6, 2017.

[6] C. Li, J. Bohren, E. Carlson, and G. D. Hager. Hierarchical semantic parsing for object pose estimation in densely cluttered scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5068–5075, 2016.

[7] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *ICCV Workshops*, pages 0–0, 2019.

[8] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019.

[9] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*. IEEE, June 2011.

[10] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018.

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2021.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012.

[15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[16] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1817–1824, 2011.

[17] Markus Suchi, Timothy Patten, and Markus Vincze. EasyLabel: A Semi-Automatic Pixel-wise Object Annotation Tool for Creating Robotic RGB-D Datasets. *arXiv:1902.01626*, 2019.

[18] A. Aakerberg, K. Nasrollahi, and T. Heder. Improving a deep learning based rgb-d object recognition model by ensemble learning. In *IPTA*, pages 1–6, 2017.

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

[20] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[21] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

[23] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[25] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[26] Dima Damen, Will Price, Evangelos Kazakos, Antonino Furnari, and Giovanni Maria Farinella. Epic-kitchens - 2019 challenges report. https://epic-kitchens.github.io/Reports/EPIC-Kitchens-Challenges-2019-Report.pdf, 2019.

[27] Dima Damen, Evangelos Kazakos, Will Price, Jian Ma, and Hazel Doughty. Epic-kitchens-55 - 2020 challenges report. https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2020-Report.pdf, 2020.

[28] Iulia-Alexandra Lungu, Federico Corradi, and Tobi Delbrück. Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1. IEEE, 2017.

[29] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10:405, 2016.

[30] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in neurorobotics*, 13:38, 2019.

[31] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017.

[32] Ajay Vasudevan, Pablo Negri, Bernabe Linares-Barranco, and Teresa Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 675–682. IEEE, 2020.

[33] F. M. Carlucci, P. Russo, and B. Caputo. (DE)^{2}CO: Deep depth colorization. *IEEE Robotics and Automation Letter (RA-L)*, 3(3):2386–2396, 2018.

[34] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794*, 2018.

[35] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. LSTA: Long Short-Term Attention for Egocentric Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[36] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[37] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020.

[38] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[40] PAUL BERTELSON and BÉATRICE DE GELDER. The psychology of multimodal perception. In *Crossmodal Space and Crossmodal Attention*, pages 141–177. Oxford University Press, April 2004.

[41] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. 2023.

[42] F. M. Carlucci, P. Russo, and B. Caputo. (DE)^{2}CO: Deep depth colorization. *RA-L*, 3(3):2386–2396, 2018.

[43] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[44] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018.

[45] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9954–9963, 2019.

[46] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[47] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[48] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.

[49] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[50] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.

[51] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021.

[52] Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *arXiv preprint arXiv:2209.08691*, 2022.

[53] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[54] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. *arXiv preprint arXiv:2110.10101*, 2021.

[55] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021.

[56] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021.

[57] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *arXiv preprint arXiv:2110.15128*, 2021.

[58] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

[59] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007.

[60] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. MARS: Motion-Augmented RGB Stream for Action Recognition. In *CVPR*, 2019.

[61] Mohammad Reza Loghmani, Mirco Planamente, Barbara Caputo, and Markus Vincze. Recurrent convolutional fusion for rgb-d object recognition. *IEEE Robotics and Automation Letters*, 4(3):2878–2885, 2019.

[62] Mirco Planamente, Chiara Plizzari, Marco Cannici, Marco Ciccone, Francesco Strada, Andrea Bottino, Matteo Matteucci, and Barbara Caputo. Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *arXiv preprint arXiv:2103.12768*, 2021.

[63] Mirco Planamente, Andrea Bottino, and Barbara Caputo. Self-supervised joint encoding of motion and appearance for first person action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8751–8758. IEEE, 2021.

[64] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1807–1818, 2022.

[65] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19935–19947, 2022.

[66] Björn Browatzki, Jan Fischer, Birgit Graf, Heinrich H Bülthoff, and Christian Wallraven. Going into depth: Evaluating 2d and 3d cues for object

classification on a new, large-scale object dataset. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1189–1195. IEEE, 2011.

[67] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 821–826, 2011.

[68] Manuel Blum, Jost Tobias Springenberg, Jan Wülfing, and Martin Riedmiller. A learned feature descriptor for object recognition in rgb-d data. In *2012 IEEE International Conference on Robotics and Automation*, pages 1298–1303. IEEE, 2012.

[69] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Unsupervised feature learning for rgb-d based object recognition. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, pages 387–402. Springer, 2013.

[70] Umar Asif, Mohammed Bennamoun, and Ferdous Sohel. Efficient rgb-d object categorization using cascaded ensembles of randomized decision trees. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1295–1302. IEEE, 2015.

[71] Max Schwarz, Hannes Schulz, and Sven Behnke. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1329–1335. IEEE, 2015.

[72] A. Eitel, T. Springenberg, L. Spinello M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *IROS*, pages 681–687, 2015.

[73] Anran Wang, Jiwen Lu, Jianfei Cai, Tat-Jen Cham, and Gang Wang. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Transactions on Multimedia*, 17(11):1887–1898, 2015.

[74] Ziyan Wang, Jiwen Lu, Ruogu Lin, Jianjiang Feng, et al. Correlated and individual multi-modal deep learning for rgb-d object recognition. *arXiv preprint arXiv:1604.01655*, 2016.

[75] Ying Zhang, Maoliang Yin, Heyong Wang, and Changchun Hua. Cross-level multi-modal features learning with transformer for rgb-d object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[76] Xianfa Xu, Zhe Chen, and Fuliang Yin. Cutresize: Improved data augmentation method for rgb-d object recognition. *IEEE Robotics and Automation Letters*, 7(1):183–190, 2021.

[77] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 345–360. Springer, 2014.

[78] Fabio Maria Carlucci, Paolo Russo, and Barbara Caputo.Θ2 co: Deep depth colorization. *IEEE Robotics and Automation Letters*, 3(3):2386–2393, 2018.

[79] Ali Caglayan, Nevrez Imamoglu, Ahmet Burak Can, and Ryosuke Nakamura. When cnns meet random rnns: Towards multi-level analysis for rgb-d object and scene recognition. *Computer Vision and Image Understanding*, 217:103373, 2022.

[80] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *Computational Visual Media*, 7:37–69, 2021.

[81] Changhyun Choi and Henrik I Christensen. 3d pose estimation of daily objects using an rgb-d camera. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3342–3349. IEEE, 2012.

[82] Chenrui Wu, Long Chen, Shenglong Wang, Han Yang, and Junjie Jiang. Geometric-aware dense matching network for 6d pose estimation of objects from rgb-d images. *Pattern Recognition*, page 109293, 2023.

[83] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015.

[84] Asmita Bagate and Medha Shah. Human activity recognition using rgb-d sensors. In *2019 international conference on intelligent computing and control systems (ICCS)*, pages 902–905. IEEE, 2019.

[85] Yao Huang and Jianyu Yang. A multi-scale descriptor for real time rgb-d hand gesture recognition. *Pattern Recognition Letters*, 144:97–104, 2021.

[86] Yunan Li, Qiguang Miao, Kuan Tian, Yingying Fan, Xin Xu, Rui Li, and Jianfeng Song. Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 25–30. IEEE, 2016.

[87] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press.

[88] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[89] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016.

[90] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019.

[91] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

[92] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[93] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1111, 2020.

[94] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[95] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016.

[96] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[97] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[98] Alexandros Stergiou and Ronald Poppe. Multi-temporal convolutions for human action recognition in videos. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.

[99] Swathikiran Sudhakaran and Oswald Lanz. Convolutional long short-term memory networks for recognizing first person interactions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[100] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift-fuse for video action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. volume 30, 2017.

[102] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

[103] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[104] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1910–1921, 2022.

[105] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. volume 45, pages 87–110. IEEE, 2022.

[106] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826, 2023.

[107] Anwaar Ulhaq, Naveed Akhtar, Ganna Pogrebna, and Ajmal Mian. Vision transformers for action recognition: A survey. 2022.

[108] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[109] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[110] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[111] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[112] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253, 2019.

[113] Tsung-Ming Tai, Giuseppe Fiameni, Cheng-Kuang Lee, Simon See, and Oswald Lanz. Unified recurrence modeling for video action anticipation. *arXiv preprint arXiv:2206.01009*, 2022.

[114] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021.

[115] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[116] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015.

[117] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[118] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[119] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[120] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[121] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[122] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12249–12256, 2020.

[123] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[124] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 154–171. Springer, 2020.

[125] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016.

[126] Jian Ma and Dima Damen. Hand-object interaction reasoning. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2022.

[127] Zeynep Gökce and Selen Pehlivan. Temporal modelling of first-person actions using hand-centric verb and object streams. *Signal Processing: Image Communication*, 99:116436, 2021.

[128] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957, 2015.

[129] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020.

[130] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020.

[131] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[132] Minlong Lu, Ze-Nian Li, Yueming Wang, and Gang Pan. Deep attention network for egocentric action recognition. *IEEE Transactions on Image Processing*, 28(8):3703–3713, 2019.

[133] Zehua Zhang, David Crandall, Michael Proulx, Sachin Talathi, and Abhishek Sharma. Can gaze inform egocentric action recognition? In *2022 Symposium on Eye Tracking Research and Applications*, pages 1–7, 2022.

[134] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[135] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*, 2022.

[136] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. Predicting gaze from egocentric social interaction videos and imu data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 717–722, 2021.

[137] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[138] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *Advances in Neural Information Processing Systems*, 35:13800–13813, 2022.

[139] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021.

[140] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, volume 2, page 3, 2014.

[141] Yao Lu and Walterio W Mayol-Cuevas. Understanding egocentric hand-object interactions from hand pose estimation. *arXiv preprint arXiv:2109.14657*, 2021.

[142] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018.

[143] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. *arXiv preprint arXiv:2106.02036*, 2021.

[144] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2016.

[145] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

[146] Yong Jae Lee and Kristen Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015.

[147] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3446–3455, 2021.

[148] Daksh Thapar, Aditya Nigam, and Chetan Arora. Anonymizing egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2320–2329, 2021.

[149] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[150] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[151] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, October 2019.

[152] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217. JMLR. org, 2017.

[153] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017.

[154] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[155] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[156] Paolo Russo, Fabio M. Carlucci, Tatiana Tommasi, and Barbara Caputo. From source to target and back: symmetric bi-directional adaptive gan. In *CVPR*, 2018.

[157] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018.

[158] Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953, 2019.

[159] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *AAAI*, pages 5940–5947, 2020.

[160] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[161] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3801–3809, 2018.

[162] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[163] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[164] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4500–4509, 2018.

[165] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8503–8512, 2018.

[166] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[167] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

[168] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[169] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[170] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. 2017.

[171] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR, 2017.

[172] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE international conference on computer vision*, pages 5067–5075, 2017.

[173] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[174] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain Separation Networks. In *NeurIPS*, 2016.

[175] J. Xu, L. Xiao, and A. M. López. Self-supervised domain adaptation for computer vision tasks. volume 7, pages 156694–156706, 2019.

[176] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[177] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[178] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels. *arXiv preprint arXiv:2003.08264*, 2020.

[179] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019.

[180] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297*, 2020.

[181] Jianhua Guo, Jingyu Yang, Huanjing Yue, and Kun Li. Unsupervised domain adaptation for cloud detection based on grouped features alignment and entropy minimization. volume 60, pages 1–13. IEEE, 2021.

[182] Yuan Liqiang, Marius Erdt, and Lipo Wang. Source protection domain adaptation by gumbel-min-max entropy minimization.

[183] Xiaofu Wu, Suofei Zhang, Quan Zhou, Zhen Yang, Chunming Zhao, and Longin Jan Latecki. Entropy minimization versus diversity maximization for domain adaptation. IEEE, 2021.

[184] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[185] Pei Wang, Yun Yang, Yuelong Xia, Kun Wang, Xingyi Zhang, and Song Wang. Information maximizing adaptation network with label distribution priors for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022.

[186] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 464–480. Springer, 2020.

[187] Nakul Agarwal, Yi-Ting Chen, Behzad Dariush, and Ming-Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. *arXiv preprint arXiv:2010.09211*, 2020.

[188] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.

[189] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6321–6330, 2019.

[190] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020.

[191] A. Jamal, Vinay P. Namboodiri, Dipti Deodhare, and K. Venkatesh. Deep domain adaptation in action space. In *BMVC*, 2018.

[192] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020.

[193] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9787–9795, June 2021.

[194] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 678–695. Springer, 2020.

[195] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pages 5334–5344, 2018.

[196] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[197] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. volume 32, pages 6450–6461, 2019.

[198] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018.

[199] Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio Maria Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. IEEE, 2021.

[200] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[201] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[202] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. volume 31, 2018.

[203] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 561–578. Springer, 2020.

[204] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

[205] Sarah Rastegar, Hazel Doughty, and Cees Snoek. Background no more: Action recognition across domains by causal interventions.

[206] Zhiyu Yao, Yunbo Wang, Jianmin Wang, S Yu Philip, and Mingsheng Long. Videodg: Generalizing temporal relations in videos to novel domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7989–8004, 2021.

[207] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. *arXiv preprint arXiv:2306.08713*, 2023.

[208] Jing Wang and Kuangen Zhang. Unsupervised domain adaptation learning algorithm for rgb-d staircase recognition. *arXiv preprint arXiv:1903.01212*, 2019.

[209] Kun Qian, Yanhui Duan, Chaomin Luo, Yongqiang Zhao, and Xingshuo Jing. Pixel-level domain adaptation for real-to-sim object pose estimation. *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[210] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*. ACM, October 2018.

[211] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:211–221, June 2021.

[212] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022.

[213] Sijie Hu, Fabien Bonardi, Samia Bouchafa, and Désiré Sidibé. Multi-modal unsupervised domain adaptation for semantic image segmentation. *Pattern Recognition*, page 109299, 2023.

[214] Jianming Lv, Kaijie Liu, and Shengfeng He. Differentiated learning for multimodal domain adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, October 2021.

[215] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *CVPR*, 2020.

[216] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008.

[217] P Lichtsteiner. 64x64 event-driven logarithmic temporal derivative silicon retina. In *Program 2003 IEEE Workshop on CCD and AIS*, 2003.

[218] Patrick Lichtsteiner and Tobi Delbruck. A 64x64 aer logarithmic temporal derivative silicon retina. In *Research in Microelectronics and Electronics, 2005 PhD*, volume 2, pages 202–205. IEEE, 2005.

[219] Patrick Lichtsteiner. *An AER temporal contrast vision sensor*. PhD thesis, ETH Zurich, 2006.

[220] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pages 2060–2069. IEEE, 2006.

[221] Rafael Serrano-Gotarredona, Matthias Oster, Patrick Lichtsteiner, Alejandro Linares-Barranco, Rafael Paz-Vicente, Francisco Gómez-Rodríguez, Luis Camuñas-Mesa, Raphael Berner, Manuel Rivas-Pérez, Tobi Delbruck, et al. Caviar: A 45k neuron, 5m synapse, 12g connects/s aer hardware sensory–processing–learning–actuating system for high-speed visual object recognition and tracking. *IEEE Transactions on Neural networks*, 20(9):1417–1438, 2009.

[222] Jörg Conradt, Matthew Cook, Raphael Berner, Patrick Lichtsteiner, Rodney J Douglas, and Tobi Delbruck. A pencil balancing robot using a pair of aer dynamic vision sensors. In *2009 IEEE International Symposium on Circuits and Systems*, pages 781–784. IEEE, 2009.

[223] Tobi Delbruck. Neuromorophic vision sensing and processing. In *2016 46Th european solid-state device research conference (ESSDERC)*, pages 7–14. IEEE, 2016.

[224] Shih-Chii Liu, Bodo Rueckauer, Enea Ceolini, Adrian Huber, and Tobi Delbruck. Event-driven sensing for efficient perception: Vision and audition algorithms. *IEEE Signal Processing Magazine*, 36(6):29–37, 2019.

[225] Tobi Delbruck and Manuel Lang. Robotic goalie with 3 ms reaction time at 4% cpu load using event-based dynamic vision sensor. *Frontiers in neuroscience*, 7:223, 2013.

[226] Arren Glover and Chiara Bartolozzi. Event-driven ball detection and gaze fixation in clutter. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2203–2208. IEEE, 2016.

[227] Martin Litzenberger, Bernhard Kohn, Ahmed Nabil Belbachir, Nikolaus Donath, Gerhard Gritsch, Heinrich Garn, Christoph Posch, and Stephan Schraml. Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor. In *2006 IEEE intelligent transportation systems conference*, pages 653–658. IEEE, 2006.

[228] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman. Hfirst: A temporal approach to object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2028–2040, 2015.

[229] Jun Haeng Lee, Tobi Delbruck, Michael Pfeiffer, Paul KJ Park, Chang-Woo Shin, Hyunsurk Ryu, and Byung Chang Kang. Real-time gesture interface based on event-driven processing from stereo silicon retinas. *IEEE transactions on neural networks and learning systems*, 25(12):2250–2263, 2014.

[230] Paul Rogister, Ryad Benosman, Sio-Hoi Ieng, Patrick Lichtsteiner, and Tobi Delbruck. Asynchronous event-based binocular stereo matching. volume 23, pages 347–353. IEEE, 2011.

[231] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018.

[232] Nathan Matsuda, Oliver Cossairt, and Mohit Gupta. Mc3d: Motion contrast 3d scanning. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2015.

[233] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. *IEEE transactions on neural networks and learning systems*, 25(2):407–417, 2013.

[234] Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, June 2018.

[235] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776. IEEE, 2011.

[236] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J Davison. Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43:566–576, 2008.

[237] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 349–364. Springer, 2016.

[238] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. volume 2, pages 593–600. IEEE, 2016.

[239] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. volume 3, pages 994–1001. IEEE, 2018.

[240] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019.

[241] Gregory Cohen, Saeed Afshar, Brittany Morreale, Travis Bessell, Andrew Wabnitz, Mark Rutten, and André van Schaik. Event-based sensing for space situational awareness. *The Journal of the Astronautical Sciences*, 66:125–141, 2019.

[242] Tat-Jun Chin, Samya Bagchi, Anders Eriksson, and Andre Van Schaik. Star tracking using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[243] https://github.com/uzh-rpg/event-based_vision_resources.

[244] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

[245] QingXiang Wu, Martin McGinnity, Liam Maguire, Ammar Belatreche, and Brendan Glackin. Edge detection based on spiking neural network model. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: Third International Conference on Intelligent Computing, ICIC 2007, Qingdao, China, August 21-24, 2007. Proceedings 3*, pages 26–34. Springer, 2007.

[246] Boudjelal Meftah, Olivier Lezoray, and Abdelkader Benyettou. Segmentation and edge detection based on spiking neural network model. *Neural Processing Letters*, 32:131–146, 2010.

[247] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016.

[248] Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International joint conference on neural networks (IJCNN)*, pages 1–8. ieee, 2015.

[249] János Botzheim, Takenori Obo, and Naoyuki Kubota. Human gesture recognition for robot partners by spiking neural network and classification learning. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 1954–1958. IEEE, 2012.

[250] Nitin Rathi, Priyadarshini Panda, and Kaushik Roy. Stdp-based pruning of connections and weight quantization in spiking neural networks for energy-efficient recognition. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 38(4):668–677, 2018.

[251] Yunzhe Hao, Xuhui Huang, Meng Dong, and Bo Xu. A biologically plausible supervised learning method for spiking neural networks using the symmetric stdp rule. *Neural Networks*, 121:387–395, 2020.

[252] José Antonio Pérez-Carrasco, Bo Zhao, Carmen Serrano, Begona Acha, Teresa Serrano-Gotarredona, Shouchun Chen, and Bernabé Linares-Barranco. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward convnets. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2706–2719, 2013.

[253] S Esser, P Merolla, J Arthur, A Cassidy, R Appuswamy, A Andreopoulos, D Berg, J McKinstry, T Melano, D Barch, et al. Convolutional networks for fast, energy-efficient neuromorphic computing. arxiv 2016. *arXiv preprint arXiv:1603.08270*.

[254] Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.

[255] Christian Reinbacher, Gottfried Munda, and Thomas Pock. Real-time panoramic tracking for event cameras. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2017.

[256] Guillermo Gallego, Jon EA Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza. Event-based, 6-dof camera tracking from photometric depth maps. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2402–2412, 2017.

[257] Alireza Khodamoradi and Ryan Kastner. $o(n)$ o (n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Transactions on Emerging Topics in Computing*, 9(1):15–23, 2018.

[258] Daniel Czech and Garrick Orchard. Evaluating noise filtering for event-based asynchronous change detection image sensors. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 19–24. IEEE, 2016.

[259] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V*, pages 308–324. Springer, 2019.

[260] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126:1381–1393, 2018.

[261] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. *IEEE Robotics and Automation Letters*, 4(2):816–822, 2019.

[262] Rafael Serrano-Gotarredona, Teresa Serrano-Gotarredona, Antonio Acosta-Jiménez, Clara Serrano-Gotarredona, José A Pérez-Carrasco, Bernabé Linares-Barranco, Alejandro Linares-Barranco, Gabriel Jiménez-Moreno, and Antón Civit-Ballcels. On real-time aer 2-d convolutions hardware for neuromorphic spike-based cortical processing. *IEEE Transactions on Neural Networks*, 19(7):1196–1219, 2008.

[263] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

[264] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018.

[265] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[266] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 491–501, 2019.

[267] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020.

[268] Junming Chen, Jingjing Meng, Xinchao Wang, and Junsong Yuan. Dynamic graph cnn for event-camera based gesture recognition. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020.

[269] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019.

[270] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432. IEEE, 2021.

[271] Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020.

[272] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 235–251, 2018.

[273] Sio-Hoi Ieng, Joao Carneiro, Marc Osswald, and Ryad Benosman. Neuromorphic event-based generalized time-based stereovision. *Frontiers in neuroscience*, 12:442, 2018.

[274] Zhen Xie, Shengyong Chen, and Garrick Orchard. Event-based stereo depth estimation using belief propagation. *Frontiers in neuroscience*, 11:535, 2017.

[275] Jianhao Jiao, Huaiyang Huang, Liang Li, Zhijian He, Yilong Zhu, and Ming Liu. Comparing representations in tracking for event camera-based slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1369–1376, 2021.

[276] Ignacio Alzugaray and Margarita Chli. Asynchronous corner detection and tracking for event cameras in real time. *IEEE Robotics and Automation Letters*, 3(4):3177–3184, 2018.

[277] Elias Mueggler, Chiara Bartolozzi, and Davide Scaramuzza. Fast event-based corner detection. 2017.

[278] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[279] Yusuke Sekikawa, Kosuke Hara, and Hideo Saito. Eventnet: Asynchronous recursive event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2019.

[280] Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. Space-time event clouds for gesture recognition: From rgb cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1826–1835. IEEE, 2019.

[281] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision*, pages 136–152. Springer, 2020.

[282] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.

[283] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[284] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019.

[285] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5831–5838. IEEE, 2020.

[286] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019.

[287] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020.

[288] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. volume 43, pages 1964–1980. IEEE, 2019.

[289] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018.

[290] Biyin Wang, Xiaojin Zhao, Yue Zheng, and Chip-Hong Chang. An in-pixel gain amplifier based event-driven physical unclonable function for cmos

dynamic vision sensors. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019.

[291] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020.

[292] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021.

[293] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019.

[294] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020.

[295] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Proc. Eur. Conf. Comput. Vis.*, pages 534–549. Springer, 2020.

[296] Xun Xiao, Xiaofan Chen, Ziyang Kang, Shasha Guo, and Lei Wang. A spatio-temporal event data augmentation method for dynamic vision sensor. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part VI*, pages 422–433. Springer, 2023.

[297] Fuqiang Gu, Weicong Sng, Xuke Hu, and Fangwen Yu. Eventdrop: Data augmentation for event-based learning. *arXiv preprint arXiv:2106.05836*, 2021.

[298] Guobin Shen, Dongcheng Zhao, and Yi Zeng. Eventmix: An efficient data augmentation strategy for event-based learning. *Information Sciences*, page 119170, 2023.

[299] Yuhang Li, Youngeun Kim, Hyoungseob Park, Tamar Geller, and Priyadarshini Panda. Neuromorphic data augmentation for training spiking neural networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 631–649. Springer, 2022.

[300] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2022.

[301] Dayuan Jian and Mohammad Rostami. Unsupervised domain adaptation for training event-based networks using contrastive learning and uncorrelated conditioning. *arXiv preprint arXiv:2303.12424*, 2023.

[302] M.R. Loghmani, B. Caputo, and M. Vincze. Recognizing objects in-the-wild: Where do we stand? In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2170–2177, 2018.

[303] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, 2015.

[304] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[305] Marco Cannici, Chiara Plizzari, Mirco Planamente, Marco Ciccone, Andrea Bottino, Barbara Caputo, and Matteo Matteucci. N-rod: a neuromorphic dataset for synthetic-to-real domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2021.

[306] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 3889–3897, 2015.

[307] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981, 2013.

[308] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[309] A. Eitel, T. Springenberg, L. Spinello M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, 2015.

[310] F. Carlucci, P. Russo, and B. Caputo. A deep representation for depth images from synthetic data. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1362–1369, 2017.

[311] J. Chung, Ç. Gülçehre, K. Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[312] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014: 13th European Conference, Proceedings, Part I*, pages 818–833, 2014.

[313] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

[314] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[315] J. Schmidhuber and S. Heil. Sequential neural text compression. *IEEE Transactions on Neural Networks*, 7(1):142–146, 1996.

[316] M.V. Mahoney. Fast text compression with neural networks. In *Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 230–234, 2000.

[317] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[318] Pretrained resnet-18. https://github.com/HolmesShuan/ResNet-18-Caffemodel-on-ImageNet. Accessed: 21-04-2018.

[319] A. Wang, J. Lu, J. Cai, T. J. Cham, and G. Wang. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Transactions on Multimedia*, 17(11):1887–1898, 2015.

[320] W. Li, Z. Cao, Y. Xiao, and Z. Fang. Hybrid rgb-d object recognition using convolutional neural network and fisher vector. In *Chinese Automation Congress (CAC)*, pages 506–511, 2015.

[321] Z. Wang, R. Lin, J. Lu, J. Feng, and J. Zhou. Correlated and individual multimodal deep learning for RGB-D object recognition. *CoRR*, abs/1604.01655, 2016.

[322] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[323] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998, 2018.

[324] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3D scene labeling. In *ICRA*, pages 3050–3057, 2014.

[325] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015.

[326] Li Yi, Lin Shao, Manolis Savva, Haibin Huang, Yang Zhou, Qirui Wang, Benjamin Graham, Martin Engelcke, Roman Klokov, Victor Lempitsky, et al. Large-scale 3d shape reconstruction and segmentation from shapenet core55. *arXiv preprint arXiv:1710.06104*, 2017.

[327] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2012.

[328] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[329] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *NeurIPS*, pages 9758–9769, 2018.

[330] J. Deng, W. Dongand R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[331] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[332] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

[333] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019.

[334] Yves Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Adv. Neural Inform. Process. Syst.*, volume 367, pages 281–296, 01 2004.

[335] Blender. Accessed: Feb. 24, 2021.

[336] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

[337] Xu Jiaolong, Xiao Liang, and Antonio M. López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.

[338] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[339] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[340] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[341] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[342] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[343] Deepak Pathak, Ross B. Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6024–6033, 2017.

[344] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2443–2451, 2015.

[345] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[346] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, 2019.

[347] Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng. Adding attentiveness to the neurons in recurrent neural networks. *CoRR*, abs/1807.04445, 2018.

[348] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019.

[349] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 20–36, Cham, 2016. Springer International Publishing.

[350] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4498–4507, 2016.

[351] Xuan Son Nguyen, Luc Brun, Olivier Lézoray, and Sébastien Bougleux. A neural network based on spd manifold learning for skeleton-based hand gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12036–12045, 2019.

[352] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Baidu-uts submission to the epic-kitchens action recognition challenge 2019. *CoRR*, abs/1906.09383, 2019.

[353] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019.

[354] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Fbk-hupba submission to the epic-kitchens 2019 action recognition challenge, 2019.

[355] Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn?, 2019.

[356] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[357] Nallapaneni Manoj Kumar, Neeraj Kumar Singh, and VK Peddiny. Wearable smart glass: Features, applications, current progress and challenges. In *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 577–582. IEEE, 2018.

[358] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021.

[359] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

[360] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.

[361] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[362] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*, 2018.

[363] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[364] Francesco Barbato, Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Latent space regularization for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2835–2845, June 2021.

[365] Qiang Zhou, Wen'an Zhou, Shirui Wang, and Ying Xing. Unsupervised domain adaptation with adversarial distribution adaptation network. volume 33, pages 7709–7721. Springer Science and Business Media LLC, November 2020.

[366] Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5089–5097, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[367] R. Xu, G. Li, J. Yang, and L. Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1426–1435, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society.

[368] Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018.

[369] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

[370] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[371] Jianbo Ye, Xin Lu, Zhe Lin, and James Z Wang. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. *ICLR*, 2018.

[372] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

[373] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation, July 2019.

[374] Pengfei Wei, Lingdong Kong, Xinghua Qu, Xiang Yin, Zhiqiang Xu, Jing Jiang, and Zejun Ma. Unsupervised video domain adaptation: A disentanglement perspective, 2022.

[375] John Bridle, Anthony Heading, and David MacKay. Unsupervised classifiers, mutual information and 'phantom targets. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

[376] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. Cogbeacon: A multi-modal dataset and data-collection platform for modeling cognitive fatigue. *Technologies*, 7(2):46, 2019.

[377] Chiara Plizzari, Mirco Planamente, Emanuele Alberti, and Barbara Caputo. Polito-iit submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition. 2021.

[378] Mirco Planamente, Gabriele Goletto, Gabriele Trivigno, Giuseppe Averta, and Barbara Caputo. Polito-iit-cini submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition. 2022.

[379] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. The kinetics human action video dataset. 2017.

[380] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015.

[381] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[382] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021.

[383] Xiaofu Wu, Suofei Zhang, Quan Zhou, Zhen Yang, Chunming Zhao, and Longin Jan Latecki. Entropy minimization versus diversity maximization for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021.

[384] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.

[385] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. volume 106, page 249–259, GBR, oct 2018. Elsevier Science Ltd.

[386] Yandong Guo and Lei Zhang. One-shot face recognition by promoting under-represented classes, 2017.

[387] Byungju Kim and Junmo Kim. Adjusting decision boundary for class imbalanced learning. volume 8, pages 81674–81685, 2020.

[388] Yue Wu, Hongfu Liu, Jun Li, and Yun Fu. Deep face recognition with center invariant loss. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Thematic Workshops '17, page 408–414, New York, NY, USA, 2017. Association for Computing Machinery.

[389] Mengke Li, Yiu-Ming Cheung, and Juyong Jiang. Feature-balanced loss for long-tailed visual recognition. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[390] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. 2019.

[391] Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020.

[392] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

[393] Yihan Lin, Wei Ding, Shaohua Qiang, Lei Deng, and Guoqi Li. Es-imagenet: A million event-stream classification dataset for spiking neural networks. *arXiv preprint arXiv:2110.12211*, 2021.

[394] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1312–1321, 2021.

[395] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.

[396] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017.

[397] Gucan Long, Laurent Kneip, Jose M Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. Learning image matching by simply watching video. In *European Conference on Computer Vision*, pages 434–450. Springer, 2016.

[398] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017.

[399] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.

[400] Will Price and Dima Damen. An evaluation of action recognition models on epic-kitchens. *arXiv preprint arXiv:1908.00867*, 2019.

[401] Gregory Kevin Cohen. *Event-Based Feature Detection, Recognition and Classification*. Theses, Université Pierre et Marie Curie - Paris VI ; University of Western Sydney, September 2016.

[402] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[403] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. *arXiv*, 2020.

[404] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[405] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[406] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[407] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

[408] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[409] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021.

[410] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[411] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

[412] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[413] Min-Ho Song and Rolf Inge Godøy. How fast is your body motion? determining a sufficient frame rate for an optical motion tracking system using passive markers. *PloS one*, 11(3):e0150993, 2016.

[414] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.

[415] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021.

[416] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7244–7253, 2019.

[417] Yingjie Wang, Qiuyu Mao, Hanqi Zhu, Jiajun Deng, Yu Zhang, Jianmin Ji, Houqiang Li, and Yanyong Zhang. Multi-modal 3d object detection in autonomous driving: a survey. *International Journal of Computer Vision*, pages 1–31, 2023.

[418] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multimodal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[419] Tsung-Ming Tai, Giuseppe Fiameni, Cheng-Kuang Lee, Simon See, and Oswald Lanz. Inductive attention for video action anticipation. *arXiv preprint arXiv:2212.08830*, 2022.

[420] Rosario Leonardi, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario. *arXiv preprint arXiv:2306.12152*, 2023.

[421] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023.

[422] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022.

[423] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[424] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[425] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[426] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[427] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[428] Sumanth Gurram, David Chan, Andy Fang, and John Canny. Lava: Language audio vision alignment for data-efficient video pre-training. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[429] Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, and Qin Jin. Accommodating audio modality in clip for multimodal processing. *arXiv preprint arXiv:2303.06591*, 2023.

[430] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.

[431] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021.

[432] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

[433] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022.

[434] Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20208–20217, 2022.

[435] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. *arXiv preprint arXiv:2211.13916*, 2022.

[436] Mirco Plananamente, Chiara Plizzari, and Barbara Caputo. Test-time adaptation for egocentric action recognition. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*, pages 206–218. Springer, 2022.

[437] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4):1–34, 2021.

[438] Gabriele Goletto, Mirco Planamente, Barbara Caputo, and Giuseppe Averta. Bringing online egocentric action recognition into the wild. volume 8, pages 2333–2340. IEEE, 2023.

[439] Antonino Furnari and Giovanni Maria Farinella. Towards streaming egocentric action anticipation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1250–1257. IEEE, 2022.

Appendix A

Code Repositories

The contributions presented in this thesis, as well as other work published, can be found in publicly available repositories. These repositories serve as a valuable resource for accessing the code, datasets, and other materials associated with the research.

- https://github.com/MRLoghmani/rcfusion
- https://github.com/MRLoghmani/relative-rotation
- https://github.com/DA4EVENT/home
- https://n-rod-dataset.github.io/home/
- https://egocentricvision.github.io/EgocentricVision/index.html
- https://github.com/EgocentricVision/N-EPIC-Kitchens
- https://github.com/EgocentricVision/RNA-TTA
- https://github.com/EgocentricVision/EgoWild

Appendix B

Project Contributions & Computational Resources Support

This thesis is made possible through the support and contributions of various organizations and initiatives. Specifically, it is a part of the *RoboExNovo* project (ERC grant 637076) and has received support from the *VIDESEC* project of the CINI Consortium. Computational resources for this research were provided by the Italian Institute of Technology (IIT) through their HPC infrastructure, as well as by Politecnico di Torino through HPC@PoliTo. We would like to acknowledge the CINECA award (*IsC92_EVEgo, IsCa0_EgoRobot, IsC98_RNA4DA*) received under the ISCRA initiative.